CDF/DOC/COMP_UPG/PUBLIC/6640

CDF Plan and Budget for Computing in Run 2: Second Annual Edition

Version 5
August 29, 2003

Edited by
Robert M. Harris and Frank Würthwein
Fermilab Computing Division

Contributions from
Stefano Belforte, Lucio Cerrito, Glenn Cooper, Matthew Herndon, Jason Harrington, Chris Hays, Beate Heinemann, Robert Wagner, Elizabeth Sexton Kennedy, Robert Kennedy, Arthur Kreymer, Ashutosh Kotwal, Dmitri Litvintsev, Victoria Martin, Mario Martinez-Perez, Mark Neubauer, Manfred Paulini, Rick Snider, Rick St. Denis, David Tang, Al Thomas, Andreas Warburton, David Waters, Steve Wolbers, and the CDF Institutional Computing Representatives Board.

Abstract

We discuss the plan for CDF computing in run 2, including the budget requirements necessary to meet the physics goals over the next 3 fiscal years.

Contents

1 Computing Introduction

This note presents an overview of the CDF computing plan for the next three fiscal years. It is intended as documentation for the Directors Review of Run 2 Computing on September 11-12, 2003. Computing expenditures are needed to move, store and analyze increasing amounts of data expected from the Tevatron. We have written a detailed budget plan, estimating our requirements, what we plan to procure in each fiscal year, and providing estimates of how much it will cost. The long range plan is grounded in our current direction, but given the rapid pace of technological advance, and what we will learn in the course of run 2, the probability of the details being correct drop quickly as a function of time. Therefore, the reader should not consider this detailed plan and budget for the next 3 fiscal years as cast in stone. We humbly submit it as our understanding at this time.

1.1 Computing and Analysis Model

A conceptual view of the major computing elements and data-flow at CDF at FNAL is pictured in Figure 1. Although incomplete, figure 1 presents some of the main themes of CDF computing. Raw data is acquired online and is written to a *write disk cache* before being archived in a *tape robot*. The raw data in the tape robot is read by the *production farms* where it is reconstructed and the resulting *reco* data is written back to the tape robot. In both cases, there are caches that decouple the *production farm* from the tape robot. The production farms use calibration constants replicated from the online *database* to the offline database and any other replicas (all shown as one database for simplicity). The reconstructed data is read primarily by *batch CPU* via a *read disk cache*. Some of the reconstructed data, and the majority of secondary datasets from the reconstructed data,are also stored in the disk cache in "Golden Pools" with essentially infinite cache lifetimes where they are accessible by the batch CPU. The batch CPU produces secondary datasets and root *ntuples* and writes them to output disk and also the tape robot via other *write disk caches* (distinct from read disk caches). The batch CPU makes extensive use of the offline database and its replicas. The batch CPU also analyzes the ntuples on the static disk. *Interactive CPU* and *user desktops* are used to debug problems, link jobs, and send them to the batch CPU which is the workhorse of CDF analysis. The user analysis farm is exclusively batch. Users desktops can also obtain data from the tape robot via read disk caches, write them back to the tape robot via write disk caches (not shown), and transfer ntuples and results back to their desktops from the interactive and batch CPU. User desktops and interactive CPU make use of the offline DB and its replicas.

In this model physics groups are encouraged to utilize the batch CPU to produce secondary datasets and write them to static disk and the tape robot. Users are encouraged to produce ntuples on the batch CPU and transport them back to the desktop for further analysis, but also have the option of utilizing the batch facilities for subsequent re-analysis of the ntuples. Users have access from their desktops and the interactive CPUs to the datasets on the CAF output disks. The interactive CPU provides a controlled environment for debugging and job submission.

Offsite resources contribute to this picture by adding additional CPU and disk caches. However, we do not expect to be using offsite tape archiving facilities at this point. The tape robot at FNAL thus serves the role of central storage facility for all official CDF data. In contrast, we do not require a copy of user level data to be stored centrally at FNAL, nor do we require tape storage prior to general open use of the data in CDF. More details on our future vision of bluring the distinction between offsite and onsite computing are discussed in Section 9.4.1.

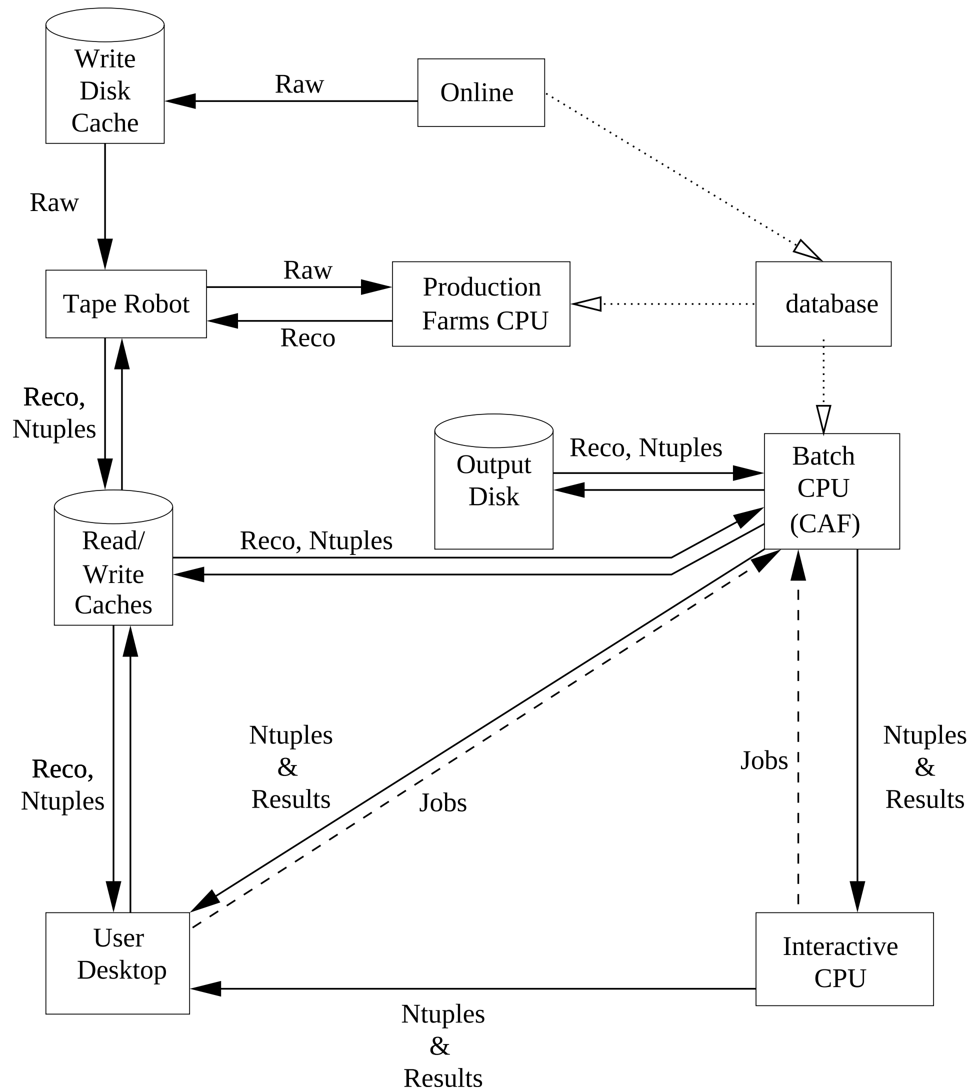


Figure 1: A simplified picture of the CDF Computing Model. Major computing elements in boxes, raw and reconstructed data flow indicated by solid lines, database constant flow indicated by dotted lines, and job flow indicated by dashed lines.Write caches are being commissioned in Fall 2003.

1.2 Requirements Model Assumptions

We have developed models of our computing requirements as a function of fiscal year. The requirements models are described in two studies [1, 2]. The old study [1] is the CDF computing plan written a year ago for the previous review. In that document the requirements models were more simplistic, and there wasn't much data available to test the models. The new study [2] builds on the previous models, updating parameters and adding important scaling behaviors, and tests the model on data from the last year. The new study has three models: a"baseline" update of the old model, and a "single-user" and "multi-user" model that introduces a new scaling behavior to the requirements. The CDF requirements we will use for our budget and procurement plan come from the "multi-user" model. We will present here some of the tables, figures and text from the new study.

CDF is planning to increase its online event logging capability. The upgrade has a significant impact on offline computing requirements [2]. The upgrade is designed to allow CDF to avoid deadtime at high luminosities and to maximize the physics program of the Tevatron by writing additional data that will increase the precision of many measurements. One particular measurement driving the upgrade, B_S mixing, is one of the most challenging and important that CDF is expected to make. It is an essential ingredient in the worldwide effort of searching for new physics by overconstraining the CKM sector of the Standard Model. The upgrade includes two changes that will increase the event logging rate from the detector: implementing raw data compression in Level-3, which should result in an increase of about 50% in the event logging rate, and an upgrade to the Consumer Server Logger (CSL) bandwidth from 20 MB/sec to 40 MB/sec. Both the baseline and upgrade paths anticipate an increase to 60 MB/sec in FY06 following previously scheduled DAQ upgrades. These changes and the fiscal years in which we assume they are implemented are shown in Table 1.

Table 1: The assumed changes to the CSL bandwidth and raw data event size, and the resulting peak event logging rate as a function of fiscal year in the baseline and upgrade scenarios. Raw data compression reduces the event size from 220 KB to 135 KB. The observed peak rates are typically about 90% of the theoretical maximum value.

	Fiscal year	02	03	04	05	06	07	08	09
Baseline	CSL rate (MB/sec)	20	20	20	20	60	60	60	60
	Raw data compression	No	No	No	No	No	No	No	No
	Implied peak event rate (Hz)	80	80	80	80	120	120	120	120
Upgrade	CSL rate (MB/sec)	20	20	20	40	60	60	60	60
	Raw data compression	No	No	Yes	Yes	Yes	Yes	Yes	Yes
	Implied peak event rate (Hz)	80	80	120	240	360	360	360	360

We employ two basic approaches to arrive at estimates for the various computing resources required by the experiment. We assume the CPU requirements factorize into two types of basic analysis behaviors. The first requirement is a high P_T dataset analysis that scales with integrated luminosity, call them "dataset-A" requirements. The second requirement is for analysis of extremely large datasets, to study bottom quarks and other high statistics physics, call them "dataset-B" requirements, with requirements that scale with the total events logged to tape. Here we assume that 400 nb of Level-3 cross section goes into dataset-A, while the balance of the logging bandwidth goes into dataset-B. We then calculate the resources required to allow 200 users to analyze 5 nb of dataset-A in a single day, and 15 users to process all of dataset-B over the course of 25 days. The dataset-A assumptions are an update to what was assumed in CDF-5914 to model our requirements one year ago. The dataset-B requirements have been added to model the load on our systems caused by the additional events that will be logged in FY04 and beyond due to the decreased event size and the anticipated increase in bandwidth capabilities of the CSL.

Some of the basic assumptions common to many of the model calculations are shown in Table 2. Included are the integrated acquired luminosity, average and peak accelerator operating efficiency, CDF data logging efficiency, typical event sizes and average logging rates. The luminosity values correspond to the "design" values [3] quoted by the Beams Division. The accelerator and logging efficiencies are typical values. Dips in the machine and logging efficiencies correspond to periods of commissioning immediately following long shutdowns, time often associated with reduced operating efficiency.

The experiment rarely operates at the peak event logging rate. More typical values are 60% to 80% of the effective peak rate. We will assume that the average logging rate is about 70% of the effective peak rate from Table 1.

The (uncompressed) baseline event size is the measured average during a set of runs in 2003 at luminosities below 4×10^{31} cm^{-2} sec^{-1}. We have ignored a known luminosity dependence in the size that is expected to produce an approximately linear 40% increase in the uncompressed data size between 10^{31} cm^{-2} sec^{-1} and 10^{32} cm^{-2} sec^{-1}. The compressed event size is calculated based upon the raw data compression factor observed in run 167024 during which the trigger table with data compression and bank dropping was tested.

Table 2: Basic assumptions and operating parameters. The average event rates are typically 70% of the observed peak rate. We ignore luminosity dependence in the event size.

Fiscal year	02	03	04	05	06	07	08	09
Integ. lum. (pb^{-1})	0.08	0.3	0.68	1.35	2.24	3.77	6.14	8.56
Integ. lum. acquired (pb^{-1})	0.06	0.25	0.6	1.2	2	3.38	5.51	7.69
Avg. acc. operating effic.	0.15	0.3	0.3	0.3	0.15	0.3	0.3	0.3
Peak acc. operating effic.	0.3	0.6	0.6	0.6	0.3	0.6	0.6	0.6
Average logging effic.	0.7	0.9	0.9	0.9	0.9	0.9	0.9	0.9
Typ. event size (baseline) (kB)	220	220	220	220	220	220	220	220
Typ. event size (upgrade) (kB)	220	220	135	135	135	135	135	135
Average event rate (baseline) (Hz)	25	50	50	50	150	150	150	150
Average event rate (upgrade) (Hz)	25	50	85	170	250	250	250	250

2 Central Analysis Farm (CAF)

The work horse for CDF user analysis is a computing cluster presently consisting of 600 CPU's, adding up to a total of 1THz of CPU cycles, accessing 180TB of disk space.

The CAF implements a vertical slice of the services ultimately anticipated for the SAM-Grid. In particular it has chosen solutions for management of input and output sandboxes, monitoring at system and user level, user interaction and diagnostics, and has provided a model for sharing of computing resources.

In this report we focus on the services the CAF provides, as well as operational aspects like reliability, human resource needs, and utilization of the cluster. We conclude with some policy issues that are noteworthy as well as expected future developments.

For implementation issues we refer the reviewer to the extensive online documentation at cdfcaf.fnal.gov . In particular, the design document, the DCAF installation guide, and the CAF software documentation. Between these three documents and the CAF User Guide all aspects of the CAF are exhaustively documented except for CAF operations. We expect to write the latter as more institutions deploy and operate CAF's offsite, thus providing the necessary incentive to document day to day operations of the CAF.

2.1 CAF services

The CAF grew out of the need to maximize the amount of computing we can provide for CDF at more or less fixed cost both in terms of hardware as well as human capital to operate the system. Fiscal pressures as well as the scale of the CDF computing challenge lead to a large batch based cluster of commodity PC hardware.

A user compiles, builds, and debugs their application on their desktop anywhere in the world. To do so we provide low bandwidth access to all CDF data files from anywhere in the world interactively. They then submit their job to the CAF by declaring their binaries, as well as a shell script to run them, a directory structure that contains both, and the level of parallelization desired. The CAF user interface forms a gzipped tar archive and sends it for execution to the CAF cluster. At the CAF site as many instances of the user tar archive are submitted to the batch system as defined by the user at submission time. At execution time, the archive is unpacked, and the user's shell script is invoked with whatever input parameters declared at submission time. One of the input parameters is an integer to distinguish between different instances of the same archive. It is then up to the user to implement the details of the parallelization based on this integer.

After the user shell script terminates the CAF creates a tar archive of the user working directory on the local node in the cluster, and copies it to a location defined by the user at submission time. In principle, the output location may be anywhere in the world. In practice we provide 50 GB scratch space per user inside the CAF. This scratch space may be accessed transparently using a set of environment variables defined by the CAF for the user. The user may access their scratch space via ftp and rootd from outside the CAF, and via ftp, rsh, rcp, fcp, and rootd from inside the CAF. We refer to this as *icaf* to indicate that the intended use is as staging area for CAF output, sort of like imap for email.

The CAF is thus receiving one tar archive with the application, and sending out as many tar archives as there are instances of the user application requested at submission time. An intelligent user will thus copy or delete all files from their working directory before exiting their shell script except for log and core files that they want back.

While the CAF is fundamentally a batch based system, we were unwilling to sacrifice the core functionality provided by an interactive system. We thus implemented not only the usual batch functionality of *submit, stat, kill* but also a core set of services that allow a user to watch their jobs as if they were running on their local desktop instead of a remote cluster.

Among those are *ls, tail, top,* and *debug.* The first three allow the user to obtain information about the local environment a given instance of their job is executing on without the need to know where that environment is located. All the user needs to specify is the instance and submission ID. The debug service allows the user to attach a gdb session to their running executable. To do this the user needs to specify the unix PID in addition to section and job id. The user may look up the latter on the CAF monitoring pages. Among other details, the web based monitoring provides CPU time consumed for all processes spawned by any instance of a users job while it is running. Please feel free to browse the cdfcaf.fnal.gov web site for more information.

Once all instances of a given submission have terminated, the CAF will parse a set of CAF logfiles created for this submission, and write a summary report to be emailed to the user. The objective with this email report is to provide the user with a quick overview of how well their submission completed. The body of the report provides sufficient information for the user to determine which instances have failed, as well as the reason for failure if known. It is thus very easy for a user to go back and debug individual instances by either inspecting the core and log files they received back with the output tar archive, or by running a specific instance interactively through a debugger. The report will soon include I/O monitoring, presently deployed only on our testcaf.

We consider the CAF services to be in their final form except for minor modifications of *stat* reporting. The one remaining service that we may develop in the future is a concatenation option. It is not unusual for a user to request 1000 instances or more at submission time. The hooks for concatenation exist but we believe that progress in data handling is required, i.e. storage and management of the intermediate results, before concatenation can be implemented in a sensible fashion.

We expect data handling to mature within FY04 and may thus revisit concatenation in FY05.

2.2 CAF Operations

To date there are close to 600 registered users, out of which roughly 100 are active on a typical day. The CAF launches roughly 10000 jobs for these users per day which are based on roughly 100-400 submission per day. Of these 10k jobs a handful ($\sim 1/1500$) fail due to CAF system failures. Roughly half of those are due to faulty hardware or OS, the remainder being attributable to known bugs in the batch system or other infrastructure software. The batch system we use is FBSNG, a FNAL product for which there are no plans of further development.

In addition to this small rate of day to day failures we witness occasional widespread failures due to known single point of failures of the CDF computing infrastructure. E.g. we have had 3 occasions within the last 9 months or so when the central code server for the CAF was either overloaded, installed with a corrupt core software piece, or had hardware or OS problems. Similarly, we had one occasion of a human error which lead to an almost 24 hour shutdown of all of CDF computing. Finally, we had a day of operational turmoil when the central building UPS in FCC failed recently. In comparison to all of these incidences there were no significant unscheduled service interruptions due to data handling within the last 3 months or so.

The CDF software does not support checkpointing. The maximum wall clock time allowed for a user application on the CAF is 48 hours. Service interruptions, even scheduled ones thus necessarily lead to job failures as it is more efficient to let some jobs fail than have all jobs drain out of the system.

The support responsibility for the CAF is shared between UCSD and FNAL, with some help from INFN. The former contribute roughly 1FTE each while the latter contributes somewhat less. This does not include efforts related to procurement nor does it account for human resources expended on networking, power, and cooling, as well as general building management, which all involve significant amounts of effort. Two FTE is thus the minimal level of operations support, including project management that is required to run the CAF. Any future development efforts require additional human resources.

UCSD is responsible for overall management of the project as well as all user interactions, system level operations, and documentation. FNAL is responsible for hardware and OS maintenance and installations. UCSD furthermore assists in operations at the level of rebooting systems during off hours as well as identifying configurations issues that impact operations. UCSD personnel work very closely with the CDF operations team, a group in the CDF department of the FNAL-CD to guarantee smooth operations of the CAF.

INFN was originally involved in the batch configuration, the user level monitoring, the organization of scratch space management, and is presently getting more involved in CAF system operations, and future directions.

In addition, there are presently two development projects, a hardware database project implemented by a Finnish student, and a monitoring project implemented by a computing professional from INFN. The former is needed to improve operations and long term tracking of hardware problems. The latter provides us the ability to monitor wait times due to tape latencies as well as total data volume consumed per dataset as well as the CPU required to do so. The monitoring project is being deployed in Fall 2003, while we expect the hardware DB work to be completed later in FY04.

We expect that a better understanding of what users do on the CAF will allow us to identify inefficiencies, as well as improved planning, and thus lead to cost savings long term.

2.3 CAF Policy

CAF policy is implemented using the concepts of queues and process types (ptypes) . FBSNG allows us to specify parameters that govern the fair share algorithm between queues and allows to set quotas and time limits for process types.

We use these to create *small, medium, large* ptypes with time limits ranging from 2-24 (4-48) hours of CPU (wall clock) time. To guarantee sufficient response time for *small* we restrict *medium* and *large* to a total quota of less than the total number of CPU's available in the farm. Fair share between users is established by having a queue per user.

We allow institutions to contribute funds to the CAF. We then guarantee that their users receive first pick of idle resources up to the level of their contribution. As a result, roughly 1/3 of the CAF was funded from sources outside the FNAL computing budget.

We also allow for disk space contributions with the understanding that 50% of the contributed disk space goes to a physics, detector, or other CDF group of the owners choice while the remainder is used at the owners discretion. An example use case for this space is the staging space used by the Monte Carlo production group for staging in MC produced offsite. The disk space used for this purpose was paid for by University of Toronto. The Canadian groups have MOU responsibility for offsite MC production coordination and a production capacity of 1 Million events per day, mostly at Toronto.

	# of nodes	# of TB	# of $1k
INFN	62	14	210
UK	10	26	150
Japan	—	—	70
KEK	10	6	55
Korea	—	5	25
Geneva	2	2	15
Canada	—	2	10
Spain	—	2	10
US	33	17	168
Total	117	~ 70	~ 700

Table 3: CAF contributions by collaborating institutions. All numbers are rounded.

Table 3 lists equipment contributed to the CAF from resources other than FNAL. In some cases like Japan the contribution is a straightforward contribution to common funds that was used for the CAF. In other cases like the UK, fileserver hardware was contributed but accounted for partially as a CPU contribution upon request by the UK groups. Finally, the US contribution is the sum across more than 10 institutions. The larger contributions among them tend to be faculty startup funds. We are presently going through another procurement cycle to spend FY03 funds. Contributions during this cycle add up to about $200k (120k INFN, 30k Germany, 17k Geneva, 30k US).

Part of the rationale behind supporting these contributions is to avoid having people try to build their own clusters in the B0 trailer complex. Investing sufficient funds into B0 trailer infrastructure to make this possible was so far judged to be too costly and thus inefficient use of finite financial resources.

2.4 CAF Utilization

CAF utilization may be looked at from three perspectives, the hardware, the batch system, and the users. These three viewpoints are presented in Figures 2,3, and 4.

Figure 2 shows actual CPU utilization and aggregate I/O bandwidth. CPU utilization ranges from about 50-90%, while an aggregate I/O of 150-400MB/sec is consumed by all user applications combined. We thus process roughly 25TB of data per day on the CAF.

All compute nodes were continuously used throughout this period. In fact, there was a significant backlog of pending requests throughout this period queued up as can be seen in the top plot of Figure 3. The CPU plot is thus really an operating efficiency plot.

While we do not have a complete quantitative understanding of the CPU utilization, we suspect that tape latencies contribute significantly to it. This is demonstrated in Figure 5. In CDF dCache, a file request can be in one of four states. If the file is not in cache then it is either being restored from tape (green), or queued up for being restored from tape (red). The y-axis in Figure 5 shows the total number of file requests. It shows that around August 9th, close to 3000 files were queued up for restores. However these restores are at least in part due to deliberate manual staging of data, and we presently do not measure the actual CPU time wasted waiting for tape restores. The means to do so will become available with the latest version of the CDF software (5.1.0).

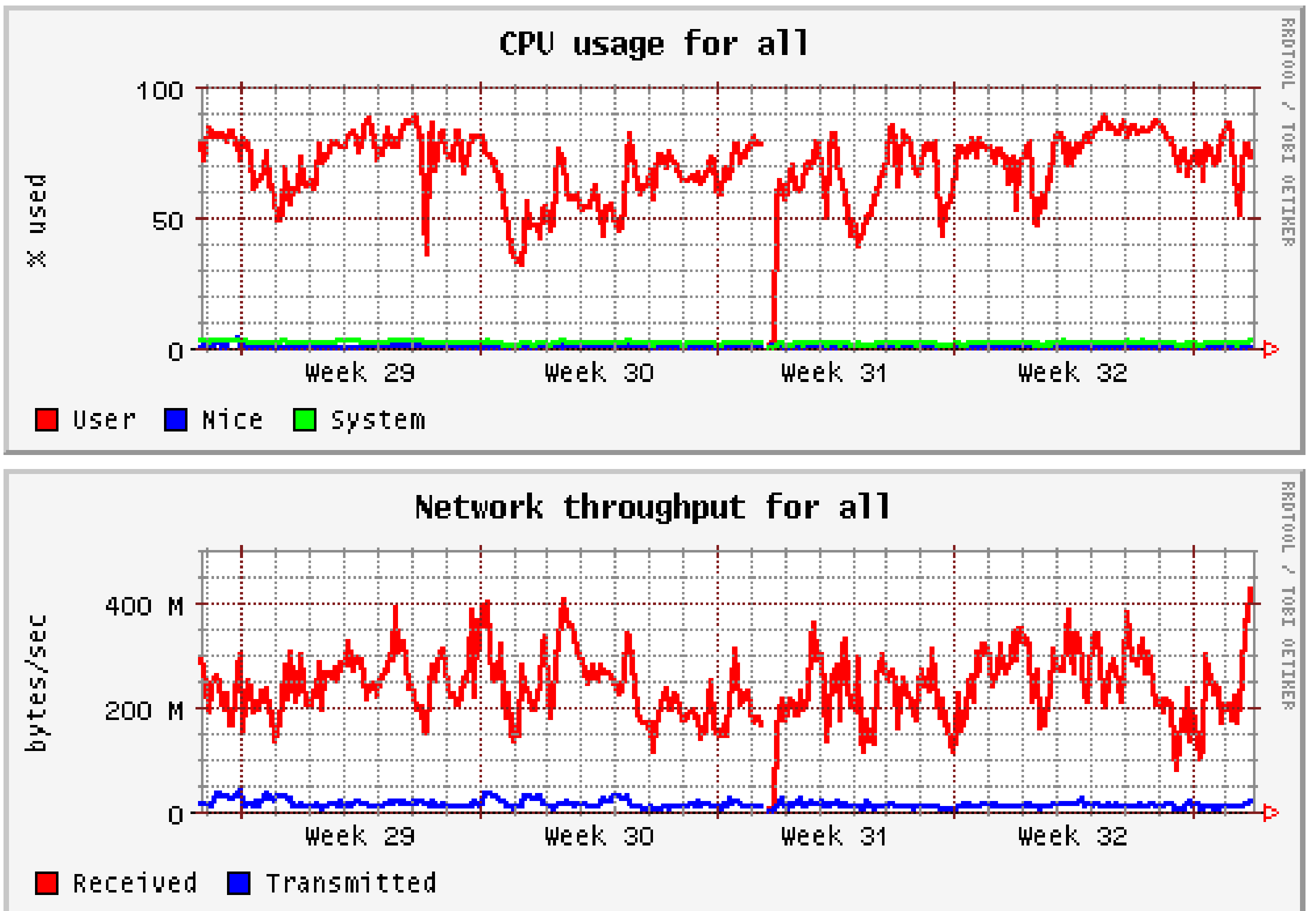


Figure 2: Average CPU utilization and aggregate I/O consumption of the CAF during July 19h to August 19th. For details see text.

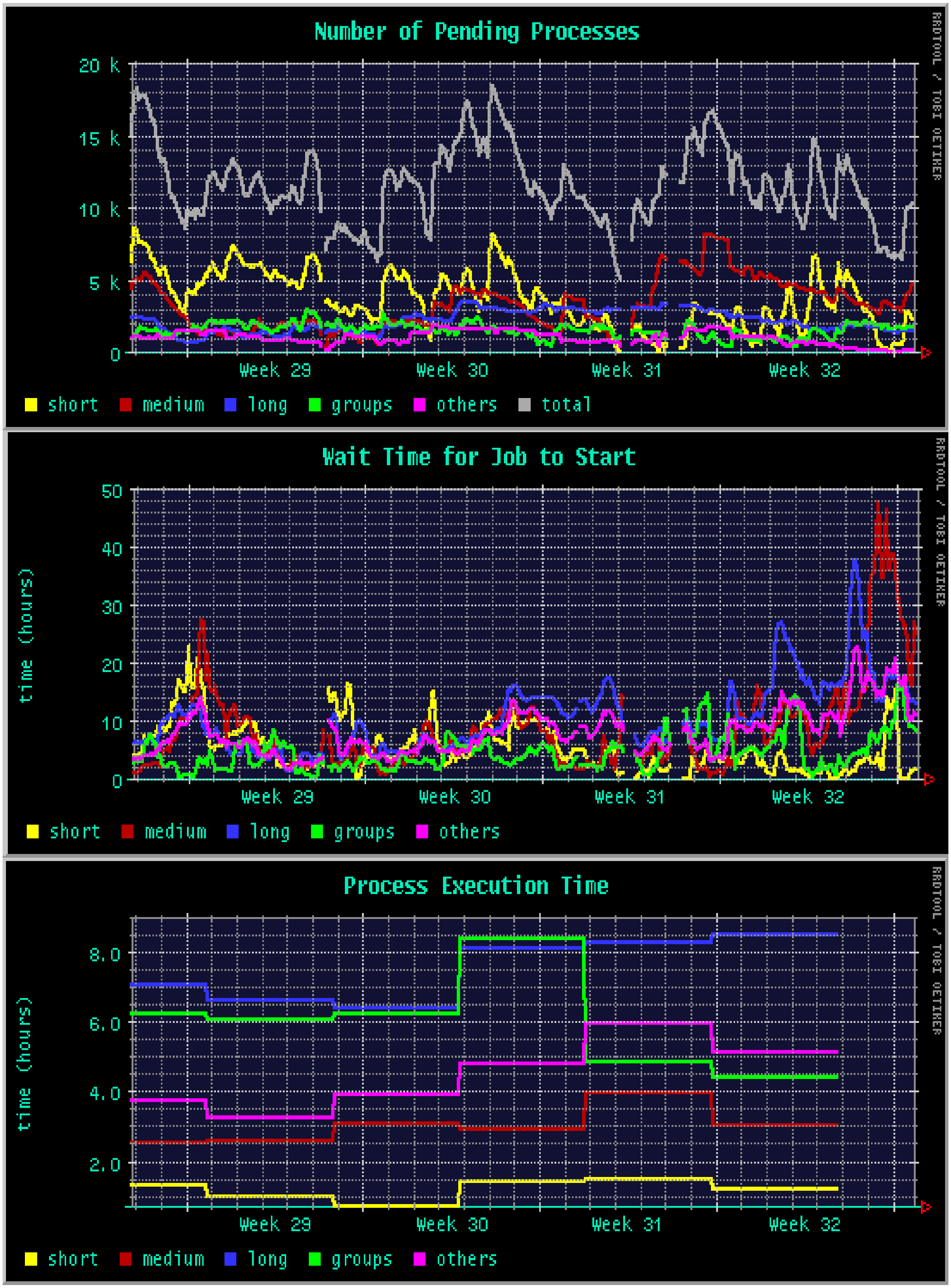


Figure 3: CAF utilization during July 18th to August 18th. For details see text.

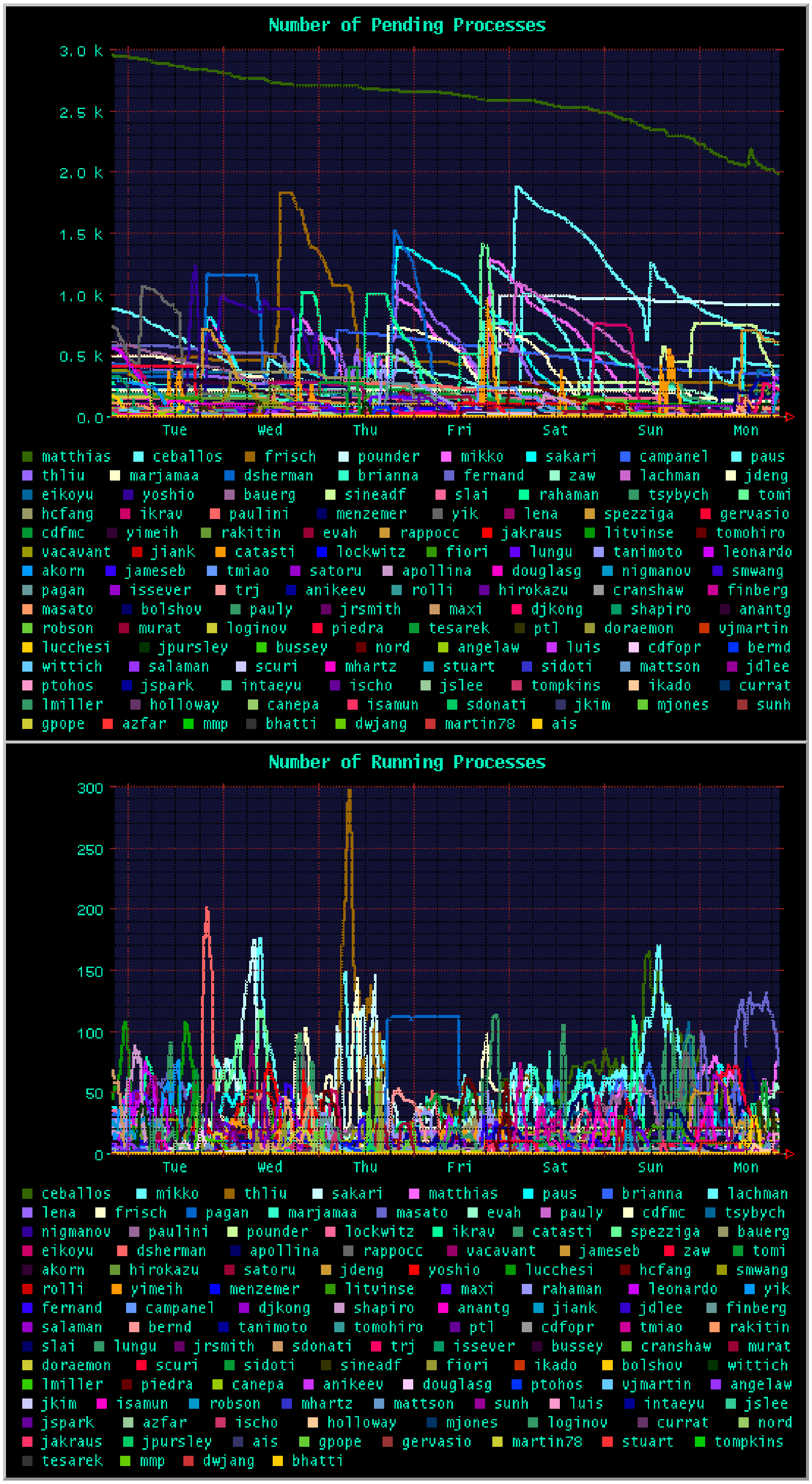


Figure 4: CAF utilization per user during August 11th to August 18th. For details see text.

If the file is in cache and the pool in which it is physically located isn't overloaded then the file is served. In dCache lingo this is called "active mover" and is indicated in yellow in Figure 5. If the pool is overloaded, then the request is queued, and accounted as "queued mover" in turquois. Figure 5 indicates that this almost never happens. Access patterns for dCache are thus responsible for only a tiny fraction of the overall inefficiency.

Similarly, hardware maintenance is responsible for less than 1% of the inefficiency. Additional possible sources of inefficiencies are startup and shutdown phases for user applications (few %), and a thread safety problem of the CDF software which occasionally leads to hung applications.

Figure 3 depicts the number of pending user applications, the time between submission and execution of the first instance, and the average execution time per instance. Color coded are the different types of processes as discussed in Section 2.3. We do not show the compute node utilization here as that is 100% all the time, thus leading to a rather boring plot.

The time period shown here is July 18th to August 18th. It thus covers part of the intense period of preparations for the Lepton-Photon 2003 (LP03) conference on August 11-16, as well as the post LP03 period. We note with some surprise that there is no easing off of use after LP03.

Figure 4 depicts pending and running jobs per user during the week of August 11th to August 18th. The colors in the two plots do not correspond to each other. Each plot's users are ordered according to the total area they occupy in the plot. Largest users come first. These plots show a variety of features which are the result of CAF policies discussed in Section 2.3. Let us discuss some of them starting with the most general.

A user tends to pend for some time before they start executing. The time they pend without executing is indicated by a flat line in the top plot. The amount of time a user pends depends on the process type they chose. Some users pend only very little or not at all because their institution owns sufficient resources in the CAF to guarantee service.

Once a user has jobs running they tend to receive significant resources and complete quickly. The exception to this rule is a user who owns resources in the CAF and keeps them fully utilized for extended periods of time, or requests vastly more work to be done than can be accomplished within a reasonable time frame. An example for the latter is user *matthias*. Further inspection of this users submitted tar archive indicates that this user is generating QCD Monte Carlo. In general, the heaviest users are either generating Monte Carlo or doing substantial reprocessing of data by themselves. These are activities that don't really require the level of data access the CAF provides and could thus be just as well done offsite. One of our goals for FY04 is to start migrating some of these activities to offsite clusters as discussed in Section 9.

As part of operations the UCSD group keeps an eye on what users do on the CAF, and interferes if it appears that someone is having trouble (many jobs core dumping), or is using the CAF inefficiently in some way. The spirit of such intervention is to help users get their job done by educating them rather than police them.

2.5 CAF future directions

We believe that the CAF's long term value lies in its services provided to the user, as well as its monitoring. The lasting intellectual value is thus in concept rather than implementation. Implementation while its cardinal weakness is also a crucial strength. It is entirely home brew with no standards other than kerberos used in its implementation. This allowed us

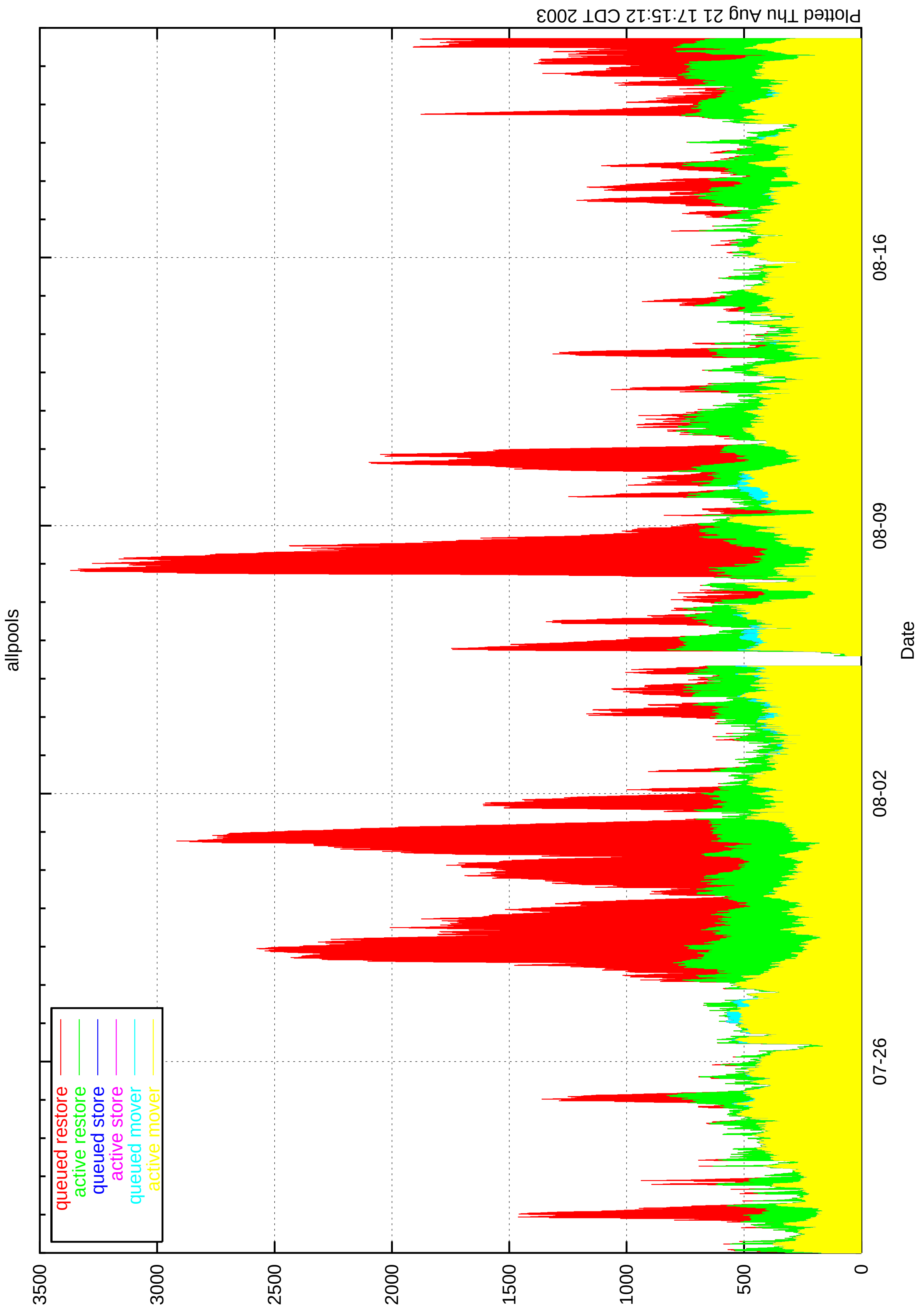


Figure 5: DCache pool queue monitoring. See text for details.

to build the first system in little more than 6 months, and will lead us to replace the present implementation as GRID standards mature and reach the same level of reliability and functionality as the CAF has today.

The challenge for the CAF is to replace part of its present implementation with emerging grid standards without sacrificing existing services or reliability. At the same time we want to extend functionality and morph the single CAF at FNAL into a cluster of CAFs across the globe that are connected via grid standards and global resource brokering in the context of SAM-grid.

At the core of our thrust to base the CAF on standards is an attempt to re-implement the CAF with Condor as underlying batch system. We expect to commit significant effort during FY04 to create CAF on Condor. This direction is also of interest to us in the context of strengthening interactive computing in CDF.

The same people from UCSD and INFN who are responsible for the CAF have joined forces with Maarten Ballentijn (MIT) to develop PEAC, Proof Enabled Analysis Center. The idea is to build a system in which batch and interactive jobs co-exist. Batch jobs consume the bulk of the CPU resources while interactive jobs suspend batch jobs whenever the former are active. The low duty cycle of interactive jobs guarantees that batch jobs aren't starved for resources.

Conceptual ideas exist, and a first prototype is planned for Supercomputing 2003 in mid November 2003. However, significant conceptual issues remain so that we consider this project to be still in its exploratory phase.

3 CAF Requirements Modeling

For the purposes of this section, the CAF consists of batch CPU and network attached disk. We assume the primary use of these machines will be production of secondary and tertiary datasets, and EDM-based user analysis. While we can reasonably estimate the CPU needed for secondary and tertiary dataset production, it is far more difficult to estimate the CPU requirements for user analysis since it depends critically upon usage patterns and analysis models, both of which are highly dependent upon the available resources, the ease of accessing those resources and the type of data upon which users rely.

In the following, we will adopt two different models to estimate the required CPU power of the CAF. As a check, we will compare the predictions of these models against the CAF utilization observed during the winter 2003 conference season. We also attempt to bound the CAF capacity needed by applying simple scaling rules to the existing CAF, which is assumed to have been adequate during the 2003 winter conferences. This exercise will also provide a feel for the overall uncertainty in the predictions.

3.1 Batch CPU model

The size of datasets in many analyses scales approximately with integrated luminosity. This behavior is particularly evident for the high P_T datasets, many of which are not pre-scaled. The user analysis model for the CAF assumes that user datasets are a fixed cross section. (Further discussion of similar analysis models can be found in Refs. [4] and [5].) We then require that the CAF provide the capacity to allow 200 simultaneous users to process 5 nb datasets in a single day assuming an event processing rate of 0.2 GHz-sec (5 events/second on a 1 GHz CPU). The latter number is based upon an informal survey of analysis jobs on

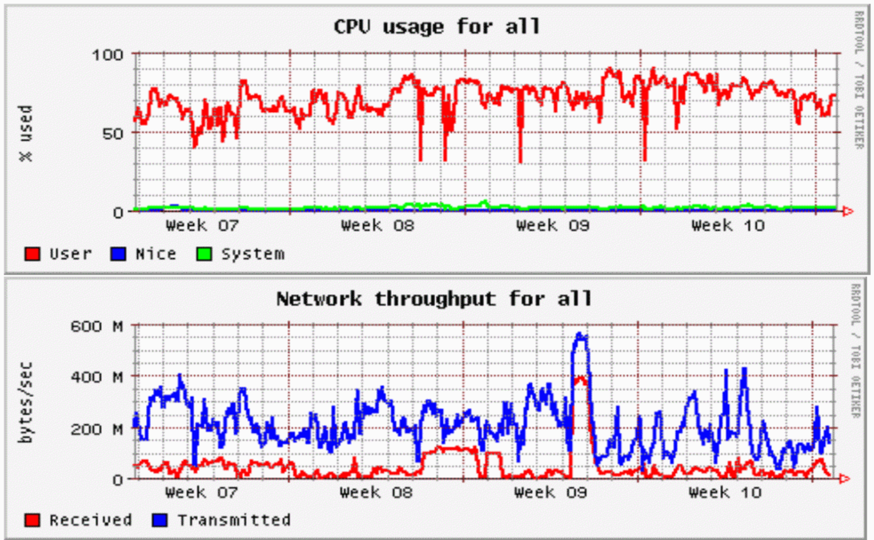


Figure 6: CPU utilization and network throughput on the CAF for a one month period during the winter conference season. The spike in the network activity was associated with tests of the data handling system.

the CAF. The observed average CPU utilization during the winter conference period was about 80% [6]. This value includes latency effects during job start-up and completion in addition to I/O, etc. We further allow an additional 30% in CPU power to arrive at the estimated needs.

To test the model, we estimate the CPU needed for the winter conferences based upon the data accumulated until that time. The model predicts a required capacity of 380 GHz; the actual size of the CAF was 939 GHz, one third of which provided exclusive access to the most popular datasets. The CPU utilization on the other two-thirds was about 50%. After removing the one third of the capacity that was unused, we find that about 630 GHZ was used during the winter conferences. The model prediction was low, but did not include secondary dataset production or Monte Carlo production or any component that scaled with the average logging rate.

Another prediction of the model is the network activity associated with event I/O. Assuming raw data event size of 180 kB, we estimate a network read rate of about 210 MB/sec. The observed rate during the relevant period is shown in Fig. 6, and is consistent with the model prediction.

It is also instructive simply to scale the existing CAF, which has proven to be adequate for current analysis needs, by either integrated luminosity or running time, and compare the result to that of the analysis-based model. Since the actual dataset has components with both types of scaling properties, we would expect that these two extrapolations represent extremes in the needed capacity of the CAF. Table 4 shows the results of this scaling exercise as a function of time. Note that in the near term, the analysis model predicts a lower CPU requirement than either naive scaling model. Since about 30% of the CAF was not utilized during the winter conferences, however, it is reasonable to conclude that both simple scaling models are over-estimates. Reducing the predictions by 30% places the analysis model predictions comfortably within the extreme values of the luminosity and run-time scaled extrapolations.

Collectively, these tests provide some confidence that the luminosity-scaled analysis model yields predictions that are within a sensible range of reality given the existing logging rates. We conclude that the luminosity-scaled model can be used to estimate the approximate cost of the CAF within the baseline model.

Table 4: Comparison of the baseline projections for the required size of the CAF (in PIII GHz equivalent) with scaling models based upon the size of the CAF used for the 2003 winter conferences. The luminosity and run-time scaling rules roughly bound the expected scaling behavior.

Fiscal year	04	05	06	07	08	09
Baseline	2240	4510	7520	12700	20730	28920
Luminosity scaled	3957	7960	13278	22420	36581	51041
Run-time scaled	2490	3485	4979	7967	10954	13942

While the luminosity-scaled model appears to capture the current computing behavior of the experiment, this is unlikely to be true for the added event logging bandwidth, since these datasets by definition grow with running time. We therefore treat the increased event logging rate separately using a run-time scaled analysis model. As before, the results will depend critically upon the analysis model we adopt. Unlike the case with the luminosity scaled models, however, no tests exist to check whether the model predictions are even within an order of magnitude.

To address this problem, we adopt two "plausible" analysis models with greatly differing computing demands. The models are plausible in that they appear to the authors to represent a reasonable mode of operation for an analysis on a large dataset, which by its nature requires more care and planning in the use of computing resources than does an analysis on smaller datasets. The first is a multi-user model similar in form to the luminosity scaled analysis model used above. We assume a total of 15 simultaneous users (among perhaps 60 within the B group working with these datasets), each of whom requires 25 days to process the entire dataset. All bandwidth above 400 nb at the average instantaneous luminosity is assumed to contribute to the run-time scaled dataset. The run-time scaling behavior begins in FY04 with the start of raw data compression. All other parameters are the same as before.

The second model represents a highly managed data reduction scheme more similar to the production of secondary datasets. A single user is assumed to access the dataset, with 20 days required to process the entire dataset. Such a model would allow a physics group to produce a highly compressed dataset several times a year (although it seems unlikely they would want to do so more than two or three times per year). All other parameters are the same, including the composition of the dataset.

Table 5 compares the CPU requirements from these models. In the multi-user model, a small number of people create a CPU demand that is comparable to the baseline demand from the rest of the experiment. A managed analysis policy for this large dataset could greatly reduce this demand, as evident from the single-user model. Since such a managed analysis policy currently does not exist, and since it is questionable whether it can be effectively implemented without slowing the output of CDF physics, we reject the single user model and adopt the multi-user model as a more realistic estimate of CDFs requirements.

Table 5: CAF CPU requirements for the baseline, single-user and multi-user models.

Fiscal year	04	05	06	07	08	09
Avg. inst. lum (10^{31} cm^{-2} sec^{-1})	4.0	7.1	19	16	25	26
Avg event rate (Hz)	85	170	250	250	250	250
Event rate into dataset (Hz)	69	140	180	190	150	150
Baseline CPU (GHz)	2240	4510	7520	12700	20730	28920
Extra CPU single-user (GHz)	122	374	528	857	1122	1384
Extra CPU multi-user (GHz)	1467	4482	6340	10283	13466	16605
Total CPU single-user (GHz)	2364	4884	8052	13560	21848	30303
Total CPU multi-user (GHz)	3709	8992	13863	22985	34193	45524

3.2 Network attached disk

The basic plan is to store as much processed data on disk as possible, while providing sufficient space for staging, data caching (if required), data validation and MC data storage. In addition to these uses, some disk is required to store ntuples or other highly compressed data samples coordinated by the physics groups. Here we will only address the portion associated with secondary dataset storage and production.

During the 2003 winter conference season, a total of 56 TB was dedicated to the CAF for data storage. This quantity met the goal of storing almost all heavily used secondary datasets on disk. We again consider the disk needed to store the fixed cross section datasets separately from the data associated with increased logging rates. The disk requirements for the fixed cross section datasets, the "A-datasets" for high P_T physics, come from scaling the winter conference usage of 56 TB by the number of raw events logged. We tried scaling by the integrated luminosity and rejected the result because it gives unphysically large values in later years. The disk requirement for the run-time scaled dataset, the large "B-datasets", is enough disk to hold the number of events that can be analyzed in seven days on the CAF.

The estimates for the total fileserver requirements is shown in Table 6. If we assume the baseline demand grows with integrated luminosity rather than data volume, then the FY04 estimate increases by about a factor of 1.5, increasing steadily to a factor of 4 by FY09. Note that all the disk in Table 6 is managed by the data handling system.

Table 6: Estimated fileserver requirements on the CAF for the baseline model, and the single and multi-user run-time scaling models.

Fiscal year	04	05	06	07	08	09
Baseline (TB)	210	294	420	672	924	1176
Upgrade: single-user model (TB)	293	629	869	1357	1824	2289
Upgrade: multi-user model (TB)	285	604	834	1300	1749	2197

3.3 CAF Procurement Plan

In the previous sections we discussed models of our requirements for batch CPU and network attached disk. Here we utilize the requirements from the multi-user upgrade model to form a procurement plan for FY04-FY06. We make this choice because the multi-user scenario more realistically models CDF analysis behavior, and we must insure that the required computing is in place to allow CDF to effectively analyze the large datasets that will result from the CSL upgrade. The relative costs of the different model scenarios can be found in a supporting document [2]. Table 7 and Table 9 lists the estimated requirements and the corresponding purchasing purchasing plan for Fermilab for FY04-06. Table 7 and Table 9 also lists the actual Fermilab procurements that took place, or are in process, for FY02-FY03 and the requirements from the previous plan [1] that drove those purchases. Table 8 and Table 10 lists CAF contributions by non-Fermilab institutions in FY02-FY03, and an estimate of CPU contributions by non-Fermilab institutions in FY04-FY06. To encourage offsite institutions to contribute to the CAF, CDF has not counted their contributions towards the total CDF requirements, and CDF has given remote institutions priority in the usage of the resources they contribute, as discussed in section 2.3.

For FY04-FY06 in Table 7 and Table 8 we assume that a dual processor costs $2.2K, which is the effective cost of a processor with speed equivalent to a 2.2 GHz PIII CPU in the middle of FY03 (equivalent to the actual $/PIII GHZ that occurred at the end of FY03 for a 2.66 GHz PIV expected to cost $1.8K). The processor speeds for FY04-FY06 then come from a Moore's law doubling in processor speed every 18 months.

For FY04-FY06 in Table 9 and Table 10 we assume that a fileserver costs $20K, which is the effective cost of a fileserver with 5 TB of disk in the middle of FY03 (equivalent to the actual $/TB that occurred at the end of FY03 for a 5 TB disk expected to cost $16.5K). The fileserver capacities for FY04-FY06 then come from a Moore's law doubling in fileserver disk size every 18 months.

From Table 7 and Table 9 we see that in FY02 the first two stages of the CAF provided the CPU and disk we estimated was needed and was available for users in January 2003, in time for winter conference analysis. Due to a smaller than anticipated Fermilab budget in FY03, the FY03 procurements for CPU fell 15% short of our estimated 2003 requirements, and they should be available in January 2004 in time for winter conference analysis. The disk procurements were on target. In FY04 we hope to move up the procurement schedule to make the resources available within that fiscal year. In FY05 we will retire all the CPU and disk purchased in FY02, and in FY06 we retire all the CPU and disk purchased in FY03. The contributions of remote institutions in Table 8 and Table 10 were significant in FY02, which allowed those contributors to complete their analysis quicker in the crunch of analysis for the Winter and Summer 2003 conferences, while their use of their own resources as opposed to FNAL resources benefited everyone. In January 2004 the stage 3 CAF will similarly have a large non-FNAL component, but in outgoing years we are anticipating a reduced contribution as the GRID makes it easier for all of CDF to take advantage of resources off-site.

Stage	FY	Needs (THz)	Duals Bought	Duals Total	Date Req.	Date Avail.	Speed (GHz)	CPU (THZ)	Total (THz)	Cost ($M)
1	02A	0.1	43	43	1/02	5/02	1.3	0.12	0.12	0.10
2	02A	0.5	136	179	8/02	1/03	1.7	0.46	0.58	0.29
3	03A	1.5	159	338	8/03	1/04	2.2	0.70	1.28	0.31
4	04E	3.7	346	674	2/04	7/04	3.5	2.42	3.70	0.76
5	05E	9.0	525-179	1030	2/05	7/05	5.6	5.88	9.00	1.16
6	06E	13.9	318-159	1189	2/06	7/06	8.8	5.60	13.90	0.70

Table 7: CAF CPU Procurements using Fermilab funds. The stage of the CAF, fiscal year,total batch CPU needed, dual CPUs bought and retired(-), dual CPUs total, the date requisitions are written, the date the CPU is available for users, speed of each CPU, incremental CPU added at that stage, total CPU available, and cost of incremental CPU to Fermilab. Numbers for FY02-FY03 are actual, numbers for FY04-FY06 are estimates and the costs are requests to meet CDF requirements.

Stage	FY	Duals Bought	Duals Total	Speed (GHz)	CPU (THZ)	Total (THz)	Fraction of CAF	Cost ($M)
1	02A	26	26	1.3	0.06	0.06	0.33	0.08
2	02A	91	117	1.7	0.31	0.37	0.39	0.18
3	03A	63	180	2.2	0.28	0.65	0.34	0.12
4	04E	30	210	3.5	0.21	0.86	0.19	0.07
5	05E	30-117	123	5.6	0.34	0.89	0.09	0.07
6	06E	30-63	90	8.8	0.53	1.14	0.08	0.07

Table 8: CAF CPU Procurements using non-Fermilab funds. The stage of the CAF, fiscal year, dual CPUs bought and retired(-), dual CPUs total, speed of each CPU, incremental CPU added at that stage, total non-FNAL CPU available, the fraction of the total CAF that this non-FNAL CPU respresents, and cost of incremental CPU to non-Fermilab institutions. Numbers for FY02-FY03 are actual, numbers for FY04-FY06 are estimates. Requisition dates and dates available for users are the same as in Table 7.

Stage	FY	Needs (TB)	Servers Bought	Servers Total	Server (TB)	Disk (TB)	Total (TB)	Cost ($M)
1	02A	30	10	10	2	20	20	0.10
2	02A	82	47	57	2	94	114	0.53
3	03A	180	18	75	5	90	184	0.34
4	04E	285	10	85	8	80	284	0.20
5	05E	604	32-57	60	13	416	606	0.64
6	06E	834	16-18	58	20	320	836	0.32

Table 9: CAF disk procurements using Fermilab funds. The stage, fiscal year, Cache disk space needs, fileservers bought and retired(-), fileservers total, the server capacity, incremental disk added at that stage, total disk available, and cost of incremental disk added. Requisition dates and dates available for users are the same as in Table 7. Numbers for FY02-FY03 are actual, numbers for FY04-FY06 are estimates and the costs are requests to meet CDF requirements.

Stage	FY	Servers Bought	Servers Total	Server (TB)	Disk (TB)	Total (TB)	Fraction of CAF Disk	Cost ($M)
1	02A	1	1	2	2	2	0.09	0.01
2	02A	34	35	2	68	70	0.38	0.34
3	03A	4	39	5	20	88	0.32	0.07
4	04E	2	41	8	16	104	0.27	0.04
5	05E	2-35	8	13	26	60	0.09	0.04
6	06E	2-4	6	20	40	80	0.09	0.04

Table 10: CAF disk procurements using non-Fermilab funds. The stage, fiscal year, fileservers bought and retired(-), fileservers total, the server capacity, incremental disk added at that stage, total disk available, the fraction of the CAF disk that this non-FNAL disk represents, and cost of incremental disk added. Requisition dates and dates available for users are the same as in Table 7.

4 Interactive Computing Systems

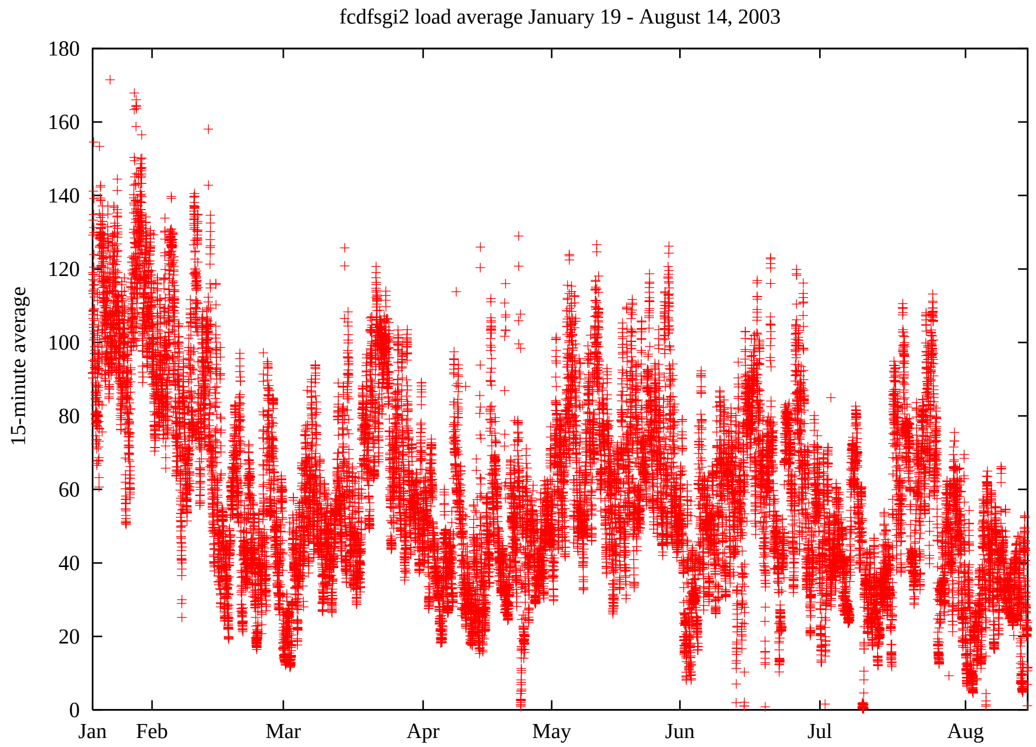


Figure 7: The load on fcdfsgi2 as a function of time during calendar 2003.

The CAF discussed in the previous section is a batch computing engine that satisfies the majority of CDFs CPU and file serving needs. The CAF is supplemented by an interactive computing system. As of August 2003, the CDF interactive computing system consisted of the legacy system fcdfsgi2, a 128 processor 300 MHz SGI SMP with roughly 45 TB of disk; fcdfhome, a NETApp serving 0.5 GB of disk per user for the home areas; and fcdfspool a NETApp serving 1 GB of disk per user for spool. Two 8-node 700 MHzIntel/Linux SMP boxes, fcdflnx2 and fcdflnx3, are used for interactive linux computing for users and offline operations. Cdfsga, a 28 processor 300 MHz SGI SMP with roughly 3 TB of disk, continued to be used for run 1 analysis.

The aging fcdfsgi2 and cdfsga SGI interactive systems cannot be maintained indefinitely, and their 300 MHz processors are increasingly avoided by CDF users. Cdfsga serves a unique role in supporting run 1 analysis on the IRIX 6.2 operating system for which run 1 code functions, and will likely need to be supported for roughly another year. Fcdfsgi2 is increasingly being used as a means of serving files stored on its common data disks to user analysis on faster Linux boxes using ROOTD. Fcdfsgi2 is thus not being used primarily as an interactive system for debugging, because developing and debugging CDF code under the IRIX operating system is inconvenient if it has to run on Linux. Fig 7 shows that the load on fcdfsgi2 has been steadily decreasing since its peak usage for the Winter 2003 conference season. The lab pays \$176K per year for the maintenance contract with SGI for all 128 processors of fcdfsgi2, and the cost scales roughly with the number of processors. If the trend in Fig. 7 continues we plan on removing 64 processors from the maintenance contract as early as January 2004 and hope to decommission the rest of the system by the end of

2004. This is warranted both for cost reasons and to discontinue support for IRIX in run2.

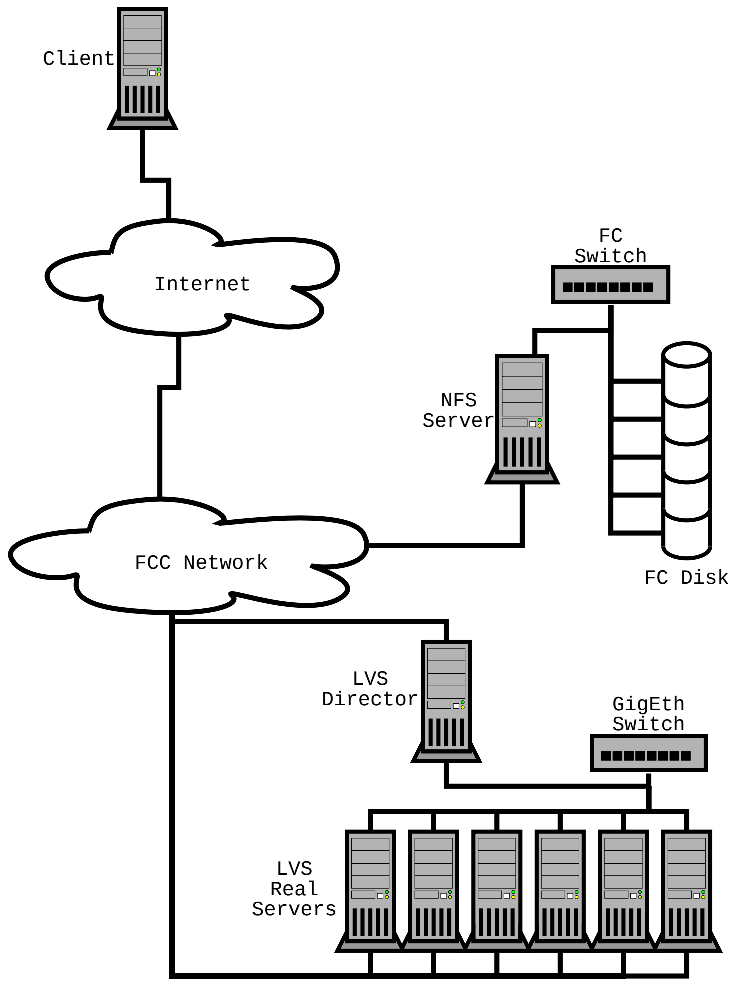


Figure 8: Schematic design of a login pool prototype to allow CDF users interactive access to a central Linux system with network attached disk (see text).

To meet our interactive computing needs, we plan during FY04 to commission and scale up a Linux interactive computing system as we decommission in phases fcdfsgi2. The Linux system is intended to provide users a "reference" platform for code development, test jobs destined for the CAF, and allow them to share files and data on disk. The Linux interactive computing system will be a login pool of CPUs with access to NFS served disk. Fig. 8 shows the prototype system which is currently being delivered and assembled. The system utilizes Linux Virtual Server to allow a client to login to anyone of six dual 3 GHz analysis nodes. A director machine, a 4-processor 2 GHz linux box, handles all requests to login to analysis nodes, and communicates with those nodes via a Gigabit Ethernet switch on a virtual private network. Data is available to the analysis nodes via a dual 3 GHz NFS server connected to 5 TB of fiber channel (FC) disk via an 8-port FC switch. The disk is recycled from the 45 TB available on fcdfsgi2. The prototype system will permit a user to interactively login to any of the six analysis nodes and run jobs that can access the NFS served disk. The prototype system will contain roughly the same amount of CPU as fcdfsgi2 and costs roughly $60K, neglecting the sunk cost of the fcdfsgi2 fiber channel disk. We hope to have the prototype burned in and ready for system testing by the end of September, with pioneer users of the alpha system sometime in October, and a beta system for the first CDF users by Thanksgiving. We hope in FY04 to increase the size of the disk pool to encompass all the fcdfsgi2 disk in stages. The additional cost of adding the remaining 40 TB of fcdfsgi2 disk to the system will be roughly $70K not including the sunk cost of the fiber channel disk. In FY05 and FY06 we will be adding CPU and disk to the system as necessary.

The largest interactive computing system at CDF are the dekstops in the CDF trailers. There are currently 360 desktops running Fermi Linux managed by two full time system administrators from the computing division. Many of the desktops are part of university cluster systems. Also in the trailers are a few specially designed clusters of PCs and disk: for example one from MIT and one from Fermilab. The current policy in the CDF trailers discourages the development of computing systems there due to power, cooling, networking and space constraints and favors instead the contribution of resources to the CDF CAF in return for priority in the queues and disk usage.

Interactive computing also includes miscellaneous expenditures not included in CAF computing. For example, in FY03 we bought a web server for $9K and a disk system for our code management node cdfpca for $8K. There will be similar miscellaneous expenditures in subsequent years. The total cost of interactive computing is estimated at $100K per year for FY04 through FY06.

5 Data Handling

From an operational and infrastructure software perspective, the DH system is comprised of user application interface (DH modules in AC++), SAM, dCache, and Enstore. One may think of this as user API, "data handling", cache management, and archival storage. Defined in this fashion, SAM's role in CDF DH is control of data movement between caches and metadata catalogue.

Operations of Enstore and dCache are largely the responsibility of the CCF and CDF departments in the computing division. SAM is a joint project between D0, FNAL-CD, and CDF. On the CDF side this is largely a responsibility of our UK collaborators, with Glasgow contributing a co-leader of the SAM joint project. The DH modules are the responsibility of CDF. They are the joint responsibility of Rutgers University and the CDF department in the FNAL-CD.

The status of DH development is such that CDF spent FY03 successfully commissioning dCache as a read cache. In the process, we overcame significant challenges in fileserver hardware, OS, and dCache software. The focus in FY04 is to put SAM into production, commission dCache as write cache, and extend the dCache-SAM combination toward user scratch space management. All of these goals are discussed in more detail below.

The remainder of this section is organized as follows. We first discuss archive related costs, as well as the model used to predict them. Costs for cache disks are discussed in Section 3.2. This is followed by a discussion of DH operations and performance. We conclude with future directions.

5.1 Data Archive Requirements

The tape archive must accommodate the raw data from the detector, the primary production datasets, secondary datasets and Monte Carlo data, all of which are assumed to be EDM-based root files. We will ignore the contributions from tertiary datasets or other

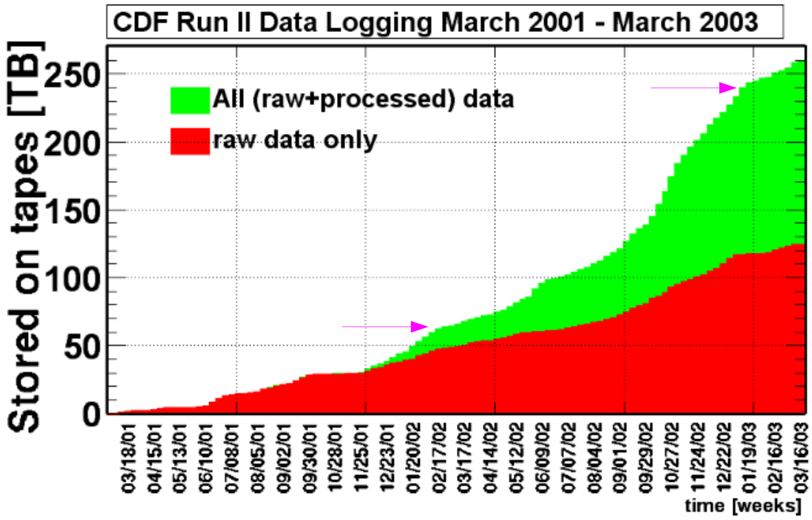


Figure 9: Volume of raw and reconstructed data stored in the tape robot as a function of time. The arrows indicate the time period during which the 2003 winter conference dataset was accumulated. The archive volume grew by about 170 TB during this time.

highly compressed files created by the physics groups, since these sources are expected to be relatively small.

To estimate the archive volume, we multiply the run-averaged logging rate by the average running efficiency within each FY (typically 0.3) and the event size after compression (220 kB for the baseline model). All logged events are assumed to go into primary and secondary datasets; the size of these events is equal to the size of production output events (about 180 kB after compression). We have ignored the dependence of the event size on the instantaneous luminosity. In practice, raw data events are expected to increase in size from about 200 kB at 1×10^{31} cm^{-2}s^{-1}, to about 280 kB at 1×10^{32} cm^{-2}s^{-1}. Both of these numbers are prior to the compression that is presently being implemented. To obtain the needed archive volume, the size is multiplied by an additional contingency factor of 1.2.

To check the calculation, we estimate the archived data volume used for the winter 2003 conferences, accumulated between February, 2002, and January, 2003, to be about 180 TB. This value follows from an average logging rate of 50 Hz over 10.4 weeks of continuous beam time (52 weeks at a 20% running efficiency). Figure 9 shows the volume of raw and reconstructed data stored in the archive as a function of time. The arrows on the plot indicate the time interval during which the winter conference dataset was collected. A total of about 170 TB was added to the archive during this time, in good agreement with our estimate.

The estimated archive volume for the baseline model is shown in Table 11 as a function of FY. For the upgrade model, we assume a raw data compression factor of 0.675 beginning in FY04, and a factor of two increase in CSL bandwidth beginning in FY05. The results for this model are shown in Table 12.

All I/O to the tape robot proceeds through a fixed number of tape drives mounted in the robot. The total I/O demands on the robot will determine the number of type of tape drives that are required. To estimate the I/O to the archive, we sum the contributions from

Table 11: Baseline archive volume and I/O rates. There are no pool-to-pool copies in the baseline model. The archive I/O estimates for the single-user model of the upgrade scenario are within a few percent of the baseline estimates.

Fiscal year	04	05	06	07	08	09
Event logging rate (Hz)	50	50	150	150	150	150
Raw data (TB)	104	104	156	311	311	311
Production output (TB)	85	85	127	254	254	254
Secondary datasets (TB)	85	85	127	254	254	254
Archive/year (TB)	273	273	410	819	819	819
Archive volume (TB)	614	887	1297	2116	2935	3754
Raw data (MB/sec)	12	12	18	36	36	36
Farms I/O (MB/sec)	43	39	48	103	120	138
CAF I/O (MB/sec)	188	339	544	907	1427	1957
Archive I/O (MB/sec)	262	408	699	1093	1638	2192

all sources: raw data logging; raw data read by the farms; production data written by the farms; production data read by the CAF; data cache misses; and secondary datasets written to the archive. We have neglected a few percent contribution from tape copies required to migrate to higher tape densities since these activities can likely be performed parasitically.

Data moving in or out of the archive is generally staged to disk first in order to adapt the I/O rate of external data consumers or producers to the I/O rate of the tape drives. This staging step implies that the archive need only provide the average read and write rates in order to keep pace with demand. To obtain the bandwidth required by raw data logging, for instance, we multiply the peak logging rate by the operating efficiency during peak periods (typically 0.6). The rate required to write output from the production farm is obtained by multiplying the raw data write rate by the ratio of production output to raw data event sizes.

At present, raw data processed on the production farm must first be written to the archive, then read back to the farm. This procedure is required due to a technical limitation of dCache. The baseline model assumes that this scheme is used throughout the entire run. Recent developments in dCache, however, will allow files to be copied from one staging pool to another, thereby avoiding the need for this read/write cycle, and one additional read of the production output by the CAF. The upgrade estimates assume that this feature is implemented starting in FY04. Finally, we assume that 10% of the file requests on the CAF result in cache misses. The resulting average load on the archive will be 10% of the read rate demand on the CAF.

We have already discussed that the data throughput estimates for the CAF during the winter conferences are in good agreement with the observed network throughput. As a further test of the I/O estimate, we can compare the total I/O load on the tape drives with the observed value during the 2003 winter conference season. Figure 10 shows the archive I/O rate during the winter and spring of 2003. The arrow points to the period of the winter conferences, during which we observed an average read rate of 6 TB/day. We estimated a demand of 56 MB/sec, or about 5 TB/day during this period.

Tables 11 and 12 present the results of these estimates as a function of FY.

Table 12: Archive volume in the upgrade scenario and multi-user model. Pool-to-pool copies begin in FY04. The requirements for FY04 and FY05 are taken from this table.

Fiscal year	04	05	06	07	08	09
Event logging rate (Hz)	85	170	250	250	250	250
Raw data (TB)	119	238	175	350	350	350
Production output (TB)	144	288	212	423	423	423
Secondary datasets (TB)	144	288	212	423	423	423
Archive/year (TB)	407	814	598	1197	1197	1197
Archive volume (TB)	748	1562	2160	3357	4553	5750
Raw data (MB/sec)	12	24	18	36	36	36
Farms I/O (MB/sec)	26	41	33	71	83	95
CAF I/O (MB/sec)	282	651	978	1597	2332	3073
Archive I/O (MB/sec)	320	716	1099	1754	2451	3204

To calculate the needed I/O capacity, we take the estimated I/O bandwidth to the archive and multiply by a contingency factor of two to take into account tape drive contention, separation of reads and writes, down-time, etc. We ignore the constraint on the total number of drives that can be used by the robots, and issues such as the mixing of drives types within a single robot. Table 13 shows the I/O requirements as a function of fiscal year for each of the three analysis models.

Table 13: Archive I/O requirements in the baseline model, the single-user analysis model and the multi-user analysis model.

Fiscal year	04	05	06	07	08	09
Baseline I/O (MB/sec)	525	909	1263	2186	3276	4385
Single-User I/O (MB/sec)	491	1093	1417	2369	3539	4727
Multi-User I/O (MB/sec)	640	1546	2059	3410	4902	6408

The archive currently contains tapes of two densities with capacities of 60 GB and 200 GB, the STK 9940A and 9940B tapes, respectively. Both densities utilize the same tape cartridge. Tapes at the lower density are therefore being copied to higher density tapes, and then added to the pool of empty high-density tapes. We assume that this process continues at a rate to meet all space demands until the supply is exhausted or a newer technology arrives, whichever comes first.

We plan to migrate to an as yet unspecified technology "X" in FY05 with twice the density of the existing 9940B tapes. The cartridges for this new technology will not be interchangeable with the current cartridges. We assume that the old tapes will be migrated to X-tapes in FY05. We have otherwise ignored total capacity constraints of the tape robot, since additional robots are relatively inexpensive.

To calculate the number of tapes needed, we take the estimated archive volume after each fiscal year and multiply it by a contingency factor of 1.2, in part to cover uncertainties

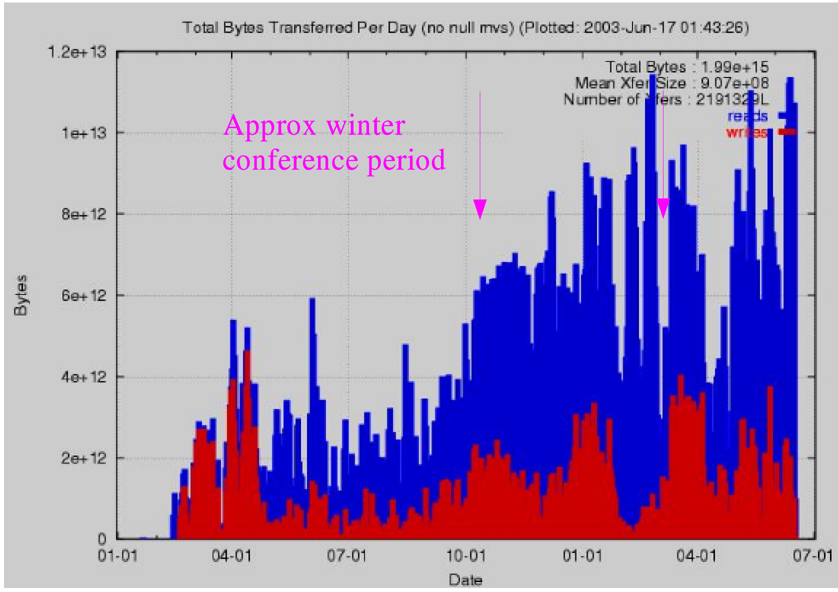


Figure 10: Tape I/O on the archive (TB/day) as a function of time. The arrows indicate the test period from Oct., 2002, through March, 2003. The model baseline model predicted a total throughput of about 5 TB/day.

in the ratio of peak to average logging rates. The requirements are shown in Table 14.

Table 14: Media requirements under the upgrade scenario. All 200 GB tapes are copied to 400 GB tapes in FY05. The "cartridges added" row includes only those tapes needed to accommodate new data, while the "migration needs" row covers the tapes required for migrating from the lower density cartridges.

Fiscal year	04	05	06	07	08	09
Capacity needed (TB)	898	1874	2592	4028	5464	6900
Tape capacity (GB)	200	400	400	400	400	400
Cartridges added	0	2443	1795	3590	3589	3590
Migration needs	0	2244	0	0	0	0

5.2 Data Archive Procurement Plan

Table 15 shows the actual drive procurements in FY02-FY03 and the planned procurements for FY04-FY06 assuming the multi-user model requirements. Procurement plans assuming different models are discussed in a supporting document [2]. The current archive uses two tape drive technologies, the STK T9940A and T9940B. These drives have maximum I/O rates of 10 MB/sec and 30 MB/sec, respectively. We are currently about half done with migrating from A to B media, and we exect that all 9940A drives will be retired in FY04.

We plan on migrating to technology "X" with an I/O capability of 60 MB/sec, twice the rate of the 9940B drives, in FY05. The 9940B drives and tape inventory will be retired in FY05 or FY06.

Stage	FY	Needs (PB)	Needs (MB/s)	Robots Total	Drives Bought	Drives Total	Date Avail	Storage (PB)	Rate (MB/s)	Cost ($M)
1	02A	0.1	70	1	10 A	10A	2/02	0.33	100	0.25
2	02A	0.2	100	2	10 B	10A+10B	2/03	1.1	300	0.52
3	03A	0.4	190	2	3 B	13B	1/04	2.2	400	0.20
4	04E	0.7	640	2	9 B	22B	7/04	2.2	660	0.27
5	05E	1.6	1550	2	21 X	11B+21X	7/05	3.3	1590	0.63
6	06E	2.2	2059	2	13 X	34X	7/06	4.4	2040	0.39

Table 15: Robot procurement plan. The stage, fiscal year stage procurement begins, total data accumulated up to the end of that fiscal year, archive I/O needs, total robots available, drives bought in that stage (A are T9940A drives and B are T9940B drives and X is a new technology), drives total after procurement, data new drives are available for use, storage capacity of robots and media in that stage, total rate available, and cost of that stage. Numbers in FY02-FY03 are actual and FY04-FY06 are estimates.

Multiplying the tape cartridge requirements in Table 14 by $75 per cartridge gives the expected media costs. Table 16 shows the actual media costs in FY02-FY03 and the expected media costs in FY04-FY06 assuming the upgrade scenario. Costs for the baseline scenario can be found in a supporting document [2].

FY	AIT-2 (PB)	9940A (PB)	9940B (PB)	X (PB)	Cost ($M)
02A	.1	.3	-	-	0.52
03A	-	.22	.24	-	0.18
04E	-	-	.90	-	0.00
05E	-	-	.90	1.9	0.35
06E	-	-	-	2.6	0.14

Table 16: Tape procurements. The fiscal year, data written to AIT-2 tapes, 9940A tapes,9940B tapes, X tapes and the total cost that FY for tape purchases. Numbers in FY02-FY03 are actual and FY04-FY06 are estimates.

5.3 Data Handling Operations and Performance

5.3.1 History

Data handling (DH) using dCache is now performing well and operations are smooth. We report here on some of the challenges we encountered on the road to stable operations.

Since Fall of 2001, CDF DH has undergone substantial changes. Throughout Winter/Spring 2001/2002 we replaced the archival storage system to a new tape, drive, and robot technology, and switched to Enstore as the archive management system. This transition was very smooth.

In Fall 2002 we started commissioning dCache as a caching layer in front of Enstore. Until then the only data access on the CAF was 15TB worth of disk space NFS mounted to all CAF worker nodes, and populated statically with the most popular datasets. The initial dCache system provided 7.5TB of cache space, and was operationally stable in late Fall. The two sets of disk space were spread across 11 of the 16 fileservers commissioned together with the CAF in May 2002. Of the remaining 5 servers, two were used for user scratch space, one for SAM development, and one for a freeware DB development project. This first generation of servers was tested in great detail for both reliability as well as performance. We found that those tests did not prepare us for running these servers in dCache. The constant write, delete, rewrite operations required in a cache system lead to severe filesystem fragmentation making the servers essentially unusable. This was fixed by switching from ext3 to xfs, and buffered to direct I/O.

In November/December 2002 we gradually commissioned 76 more fileservers and quadrupled the size of the CAF. The resulting CAF was deemed too large to reliably operate by NFS mounting all data disks. DCache was performing very well at low loads and thus went into production for all of the new CAF as planned. Under increasing load the resulting system functioned very poorly. At the peak of our problems in mid January we had a service reliability of roughly 50%, i.e. a user had a 50/50 chance of having a requested file actually delivered.

Rather than attempting to run a dysfunctional system we decided to take dCache out of production and focus on understanding and fixing our hardware and software problems. To provide service to CDF during the winter 2003 conference season we populated an additional 26TB of disk space by hand, and NFS mounted it to 200 of the new CAF CPUs. CDF thus depended on the old and the new CAF, both of which had data access only via NFS mounted disks. Access was controlled via iptable arrangements on the worker nodes.

We discovered bugs and fundamental scaling limitations in dCache, bugs in the OS, and a faulty SCSI disk in the pnfs server. Pnfs is the virtual filesystem upon which both dCache and Enstore are based. In addition, all our 1064 new WD180GB disks had a serious flaw. They would randomly drop out of the RAID system, pretending to be broken, iff load was sporadic with sufficient idle times in between. RAID systems in production were thus unstable while they were perfectly stable under constant load during burn-in and reliability testing.

CDF survived the winter conference season by supporting all 13 fileservers used to NFS mount on the 200 CPU fraction of the new CAF on a 24x7 basis, and servicing them at least twice a day by hand for disks that dropped out of RAID sets.

In parallel to this CCF, DESY, and CDF worked out the problems with dCache, and CCF helped resolve the OS problems. ASA, one of the two fileserver vendors arranged for a team of engineers from WD and 3ware to visit FNAL. With assistance from the CAF team

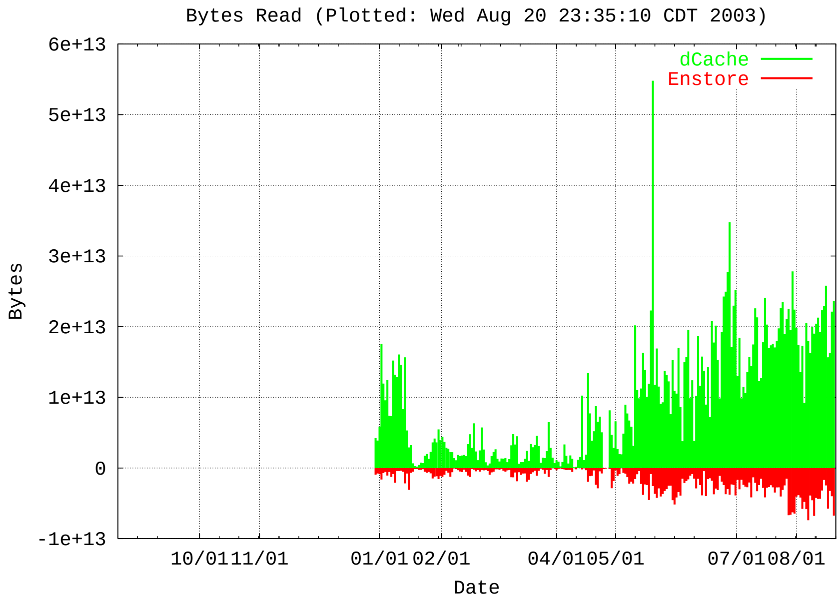


Figure 11: Number of bytes read per day from dCache. Data starts on January 1 2003 and the plot ends at late August 2003. The spike of 55 TB/day was a deliberate load test prior to declaring dCache "in production" at CDF.

WD identified and fixed the drive problems with a firmware upgrade during a second visit. The CAF team, with some assistance from a WD engineer during a third visit replaced the firmware on all 1064 drives by mid March.

As a result of an enormous amount of very hard work by a large group of people from CDF, CCF, DESY, and WD we resolved all our operational problems by middle of May 2003. The CDF dCache system has been rock solid ever since.

5.3.2 Current Status

The DH system using dCache now satisfies the I/O demand of CDF computing, a typical load of about 20 TB/day as shown in figure 11. We estimate that with the current 35 fileservers and 10 tape drives the system can provide the following average rates: 40 TB/day averaged over a week, 1.7 TB/hour averaged over a day, or 0.5 GB/sec averaged over an hour. With the same fileservers and tape drives we estimate the system can currently provide the following peak rates: 60 TB/day, 3.3 TB/hour, 1.1 GB/sec. The load test spike in Figure 11 used all 560 CAF processors simultaneously reading data from dCache and demonstrated a rate of 55 TB/day.

To maximize cache hits, thus minimizing DH related inefficiencies on the CAF, we partition the dCache system into three types of pools based on the expected access patterns:

1. "Golden Pool": secondary datasets that are most relevant for a conference season. We guarantee that those are always on disk by providing sufficient disk space to keep up with new data coming in. We arrive on the list of golden datasets in collaboration with the CDF physics coordinator.

2. "Volatile Pool": regular cache.

3. “Raw data and big buffer Pool”: some datasets, especially raw data streams, are either so large, or so infrequently used that the number of times a file is accessed while in cache is rather small. The cache thus functions more like a FIFO buffer than an actual cache.

We separate these three types of cache administratively. In addition, we administratively request users to notify DH operations group before they attempt to access large datasets that are not on the list of golden datasets. DH group then organizes prestaging for the user in an attempt to minimize waste of CPU time on the CAF due to tape latencies.

In the future, SAM will automate the prestaging because a user will declare their dataset at CAF submission time rather than runtime.

5.4 Future Directions

Crudely speaking, there are two broadly defined areas of development that remain: SAM and write caching.

5.4.1 SAM migration

There are many reasons for CDF to adopt SAM:

- Combined development and maintence of DH software with D0, thereby reducing costs by eliminating redundant solutions.
- A clear path to the future Grid-supported tools.
- DH support for offsite computing. This is discussed in detail in section 9.
- Improved operational efficiency of the CAF at FNAL, as discussed above.
- More extensive metadata catalogue functionality. Example use cases are documentation of MC production and derived datasets. At present, several physics groups produce secondary datasets which are often not very well documented.

At the time of last year’s review, CDF had just started use of SAM by some offsite users. Since then progress has been made in the following areas

- The SAM project has become a joint project with contributions from CDF and D0 and from SAM-Grid. A management structure exists with co-leadership from both experiments and a general oversite committee chaired by the computing division head.
- The SAM database schema has been modified to accomodate the needs of both experiments. In the process the experiments also learned how to cooperate and find common ground. The new schema has spawned a project to upgrade the SAM middle tier database server.
- Some users at remotes sites use SAM to do their physics analysis and are able to take advantage of their remote computing resources. Of particlar interest are Karlsruhe and Oxford. Other users have taken SAM in order to access small amounts of raw data for special studies. An example of this is Trieste.

- The SAM database has been populated with the CDF data catalog and all data in CDF are available in SAM. This was done by automatic copying of new files from DFC to SAM, and was found to be less than ideal operationally. Starting in late August 2003 we have switched to registering all files produced on the production farm automatically in both DFC and SAM at the same time.
- The CDF framework is able to switch easily to using SAM for its data handling mechanism.

There is still quite a bit of work to go before SAM is used as the predominant data handling mechanism. We will follow the mechanism that was used to bring dCache into production: run with CDF software releases that allow SAM to be chosen in an open beta testing and once it is proven that SAM can satisfy CDF needs reliably, to make SAM the default.

The crucial remaining step in the transition is to interface SAM with dCache. This is and has been for some time, the biggest hurdle to overcome in order to put SAM into production on the CAF. We are pursuing this in the context of a more general discussion on caching strategies with CCF, D0, and CDF.

The goal of this discussion is to indentify a software layering strategy within SAM to make different types of caches equivalent and pluggable, so that sites can deploy differing implementations based on their local support expertise, and site strategies. As a result, SAM will become more versatile, modular, and generally more attractive a solution for people outside of Fermilab.

This direction is very much aligned with the strategic interests of the Lab, especially in the context of grid computing.

The timeline for having a prototype implementation of SAM-dCache integration for deployment on the testcaf is November 15th. Assuming success, we will then put SAM onto the production CAF, and allow users the choice between SAM-dCache-CAF and the existing dCache-CAF combo. We strongly believe that watching users vote with their feet is the best way to verify that a technology transition was both needed, and implemented successfully.

In parallel to the migration of SAM onto the dCache-CAF, we have an offsite SAM milestone of October 15th for generating MC in Glasgow based on generator files from Enstore Patriot system at FNAL. The output will be stored back in the CDF Enstore system, and catalogued in the SAM metadata catalogue.

Depending on success, we will use this capability for at least part of our Winter 03/04 Conference MC production of roughly 150 Million events.

5.4.2 Write Caching

One of the longstanding issues with CDF computing infrastructure is the lack of a tape-less data-path. At present, a user doesn't see the data until after it has been written to and read from tape at least twice. The relatively high cost of tape drives has led CDF to acquire a relatively modest number of them. And as a result, we are relatively sensitive to tape latencies due to fluctuations in the tape access patterns.

Creating a tape-less data-path by adding write caches at the output of both CSL and the production farm will smoothen such fluctuations, and generally expedite production farm operations, and improve CAF operational efficiency. It will also simplify archiving of data by users. We expect to commission dCache write pools in Fall of 2004.

Apart from write caching in front of the robot, we also have a clear need for better support of individual user data. In a typical analysis a user starts with some secondary dataset, produced in a coordinated fashion by a physics group. The output of this processing on the CAF will generally be a quite sizable collection of relatively small output files. The user thus needs to store these files temporarily for validation, and further analysis, and possibly concatenation. In general, this processing step is done more than once in order to fix some oversight or the other. Old versions may be deleted to conserve disk space. An ideal storage system for this use case is disk resident only, and supports deletion as well as quotas.

At present, we support this activity by providing user scratch space inside the CAF. However, we know that this does not scale, especially if groups of users organize themselves to produce common datasets. As rootd is the only access mechanism to these files, and rootd doesn't support any kind of traffic shaping, we end up overloading the fileserver hardware, and crashing jobs on the CAF. This is clearly not supportable as we scale up CDF computing. We need to fully virtualize the user scratch space, and then guarantee that datasets are spread randomly across many pieces of hardware.

Furthermore, ATA disk prices per GB are presently minimal for 100-120GB disks. However, CAF operations really require only about 20GB or so data-disks. Disks this small are no longer cost effective. As a result, we tend to end up with significant amount of unused disk space across the CAF. It would be nice to fully virtualize this disk space, and make it a reliable storage system by implementing a "virtual" RAID1.

We are presently discussing an implementation of these ideas based on a SAM-dCache combo. While some of the details still need to be worked out, CDF, D0, and US-CMS have all expressed interest in this functionality, and DESY and CCF are interested in implementing it. We expect the details of the implementation to be worked out in Fall 03, and a first prototype to be ready for testing in Spring 04.

We consider this to be an important ingredient in our SAM-grid vision as described in Section 9.4.1.

6 Production Farms

6.1 History

The historical growth of the production farms is presented in table 17. As of August, 2003 the CDF production farms are 135 duals used for event processing, 16 duals used for I/O, 2 server nodes, 1 web server, consoles, and a network switch. The farms will need regular CPU increments to replace old machines and to increase CPU capacity. Some networking upgrades will be needed in future years for increased data rate.

6.2 Requirements

In table 18 we estimate the farms requirements. Starting with a peak logging rate and the running efficiency we estimate the event/year and the total events of raw data. The CPU time per event is assumed to be 5 sec/event on a 1 GHz processor, which takes into account increases in CPU time required for higher luminosity and newer versions of code. Assuming an up-time of 75%, each 1 GHz processor on the farm can process 5 million events per year. Dividing the processed events per year by 5 million gives the tabulated number of GHz required to satisfy the average rate needs of the farms. To allow the farms to keep up with

FY	Nodes (duals)	Total (duals)	Type (GHz)	CPU (GHz)	Total (GHz)
1999	50	50	PIII 0.5	50	50
2001	23	73	PIII 0.8	37	87
2001	64	137	PIII 1.0	128	215
2002	32	169	PIII 1.3	81	296
2002	32	201	AMD 1.7	107	403
2003	-50	151	PIII 0.5	-50	353
2003	64	215	PIII 2.2*	282	635
2003	-23	192	PIII 0.8	-37	598

Table 17: Farms procurement history. The fiscal year, nodes added, total nodes, speed of each node, total speed of nodes added, and total speed of all nodes in GHz. 2.2 GHz is the "PIII equivalent" speed of the PIV 2.66 GHz machines to be purchased in FY03; the actual clock speed is higher. Although the total number of duals is listed at 192 in 2003, 16 are used only for I/O, so in subsequent tables we list 176 as the number of duals in 2003, and 562 GHz the total amount of CPU in 2003.

typical data rates that occur over a few days time, a factor of 2 more CPU is required than the average CPU necessary. The factor of 2 comes from comparing the expected up-time of the system over multiple stores, roughly 60%, with the canonical uptime over a fiscal year, roughly 30%. Multiplying the average rate needs of the farms for data processing by a factor of 2 gives the data CPU needed in table 18.

6.3 Purchasing Plan

The plan for future acquisitions is presented in table 19, and for completeness we include the FY03 acquisitions that are already in process. Each FY between FY04 through FY06 we plan to purchase 64 duals and retire 64 older duals to keep the size of the farms constant while keeping the technology current. For the costs we have assumed $2.2K for each PC, $12.5K for annual network upgrades, and $37.5K for miscellaneous annual expenditures (replacement of server machines, cabling, power work, etc.). The costs do include replacement of servers but not any dCache hardware. The I/O-worker nodes are included in the number of duals replaced in FY04 and FY05. Therefore the total CPU available in FY04 and FY05 is increased by 8 nodes fewer than the 64 purchased.

6.4 Software Development and Operational Issues

The CDF farms have been extremely successful in their primary role – processing bulk data from the CDF experiment in "pseudo real" time, splitting the output into many (currently 35) output "data sets" and concatenating the output into approximately 1 Gbyte files. In addition, the farms are used to test new versions of the CDF ProductionExe (the 100K tests) and to process special runs as requested by the online system. The farm software (fps) is not static nor are CDF needs. The following are the major changes anticipated during the upcoming years.

1. "Beamline" processing prior to full reconstruction.

Fiscal Year	03	04	05	06
Peak rate (Hz)	80	120	240	360
Running Efficiency	0.2	0.3	0.3	0.3
Events/year (billions)	0.5	1.1	2.3	3.5
Total events (billions)	0.6	1.7	4.0	7.5
Re-processings	1.0	0.5	0.3	0.2
Re-processed/yr	0.6	0.9	1.2	1.5
Processed/year	1.2	2.0	3.5	5.0
Data CPU average (THz)	240	400	700	1000
Data CPU needed (THz)	480	800	1400	2000

Table 18: The farms CPU requirements. As a function of fiscal year, the peak online logging rate of raw data, the total running efficiency, the resulting raw events each year, the total raw events up to that year, the number of re-processings of the entire data sample excluding the first event reconstruction, the number of events reprocessed during that year, the total number of events reconstructed each year including all re-processings, the required CPU to process the average rate of raw data, and the total CPU needed for raw data including the ability to keep up with rates larger than average.

FY	Needs (GHz)	Nodes (duals)	Total (duals)	Type (GHz)	CPU (GHz)	Total (GHz)	Cost ($M)
2003	480	64-23	176	2.2	282	562	0.19
2004	800	64-64	176	3.5	392	842	0.19
2005	1400	64-32-32	176	5.6	627	1301	0.19
2006	2000	64-64	176	8.8	1021	2146	0.19

Table 19: Farms procurement plan. The fiscal year, total CPU needed, nodes added, total nodes, speed of each node, total speed of nodes added, and total speed of all nodes in GHz, and the cost including incidentals (see text).

The data sets coming from the farms would be more useful if the final beamline and alignment, as well as calorimetry calibration information was available along with the primary datasets. This will be implemented by running a special farmlet after raw data is taken and before the full processing occurs. The mechanics for doing this is essentially finished and will be used for the major reprocessing beginning in September 2003. The same procedure will be folded in to all processing soon. This effort is part of a larger focus on making the production output immediately useful to the end user, thus avoiding CPU expensive processing steps at the user analysis level.

2. SAM metadata use by the farms.

   The data file catalog (DFC) has been the home for all metadata for CDF data since the beginning of Run 2. We have started transitioning to the use of the SAM catalog in late August 2003 as discussed in Section 5.4.1.

3. Offline operations use of the farms.

   Up to now the CDF farms have been restricted to the ”experts”, limiting the pool of potential users and making for a rather thin coverage of the systems, especially during summer and around holidays. An important goal is to allow others, especially the CDF offline shifters, to use part or all of the farms to run 100K and other tests and eventually to run the entire farms. Part of this has already occurred and more extensive expansion of farms expertise will occur soon.

4. Multiple executables, additional output streams.

   There have been requests to provide additional functionality on the farms to allow for multiple executables (allowing the split of validation from ProductionExe) and to write some or all of the secondary data sets from the farm. Both of these require some changes to the farms control software and will be specified and implemented.

5. dCache/tape-less data paths.

   An interesting and potentially very useful enhancement to the farms operation is the ability to read and write data from/to disk (probably dCache) rather than from/to tape in Enstore, as is done today. There is no fundamental barrier to doing this, but it requires a clear plan with a design that takes into account data flow, latencies due to the production of calibration constants and the beamline positions, output data flow, and storage of data to permanent media as soon as possible to avoid loss of data. Given that the input to the farm is raw data from the experiment this change implies that the raw data logging would be written to a disk cache.

6. Offsite reprocessing.

   The farms plan given in this section assumes a roughly constant number of reprocessed events in FY05 and FY06, corresponding to only 0.2 total re-processings for the entire CDF dataset at the end of FY06. This is somewhat lower than the 0.3 reprocessing that took place for run 1. More reprocessing is possible using offsite facilities, and we expect that offsite reprocessing could be as significant as on-site reprocessing by FY06. The specific plans are still to be discussed and decided but it is clear that the use of GRID tools, SAM and other changes to the CDF offline system will allow a more transparent use of resources at Fermilab and across the world. This leads to a potential large gain in processing power, a reduction in processing latency (if large resources can be used quickly) and more flexibility and efficiency in deploying the offline production farms, the CAF, and other computing on the grid.

7. Integration into SAM-grid.

   The conceptual inverse of running reprocessing offsite is to allow other types of data processing to proceede on the production farm. This is motivated by the observation that the overal CPU utilization of the production farm is rather low by design in order to be able to keep up with data taking while data is being taken. The low duty cycle of the accelerator thus translates into a low average CPU utilization of the production farm. In the long term it would be desirable to utilize those spare production farm cycles in some fashion by integrating the farm more tighly into SAM-grid.

7 Databases

CDF currently utilizes Suns for online databases and a combination of Suns and Linux boxes for offline databases. CDF database hardware setup is listed in Table 20. The online production machine, b0dau35, and development and integration machine, b0dau36, are identical Enterprise 4500 servers with two 400 MHz processors each. The machines are dedicated to running Oracle database server version 9.2.0.3, that will be upgraded during fall 2003 shutdown to 9.2.0.4. The online production database is behind a firewall. It is divided into several applications: Trigger, Hardware, Run, Calibration and Slow Control (MCS). Write access to these applications requires running on privileged nodes located in FCC. The content of the online production database is replicated to offline production database via Oracle read-only replication. The offline production machine is fcdfora1 which is also Sun Enterprise 4500 dual 400 MHz processor server with 1.25 GB of memory and 0.5 TB of disk space. Besides replicated tables the offline production database hosts Data File Catalog and SAM [1] tables. Fcdfora2 is identical to fcdfora1 and hosts offline development and integration Oracle instances. The access to the offline production database is not restricted. Since summer 2002 the content of online production database and Data File Catalog are replicated to the more powerful machine, fcdflnx1, a quad PIII 700 MHz machine with 4 GB of memory and about 1 TB of disk space on SCSI RAID arrays.

The amount of data used by existing applications is constantly monitored. Disk space usage as a function of time is used to make projections of space needed for application in the future. Figure 12 shows DB space needs vs time for online and offline production databases. Currently the actual space used by all applications combined is about 90 GB.with, 153 GB including high water mark. Calibration and Slow control applications use the most space.

7.1 Offline Production Database

The offline production database machine fcdfora1 is using 300 GB[2] out of 500 GB of available disk, and given the design of the system it cannot be easily expanded to meet future disk and CPU needs. In less than a year the database size will exceed the capacity. The machine also has severe I/O limitations which have caused serious degradation of performance of the database. The cost of fixing these limitations is equivalent to replacing the entire system.

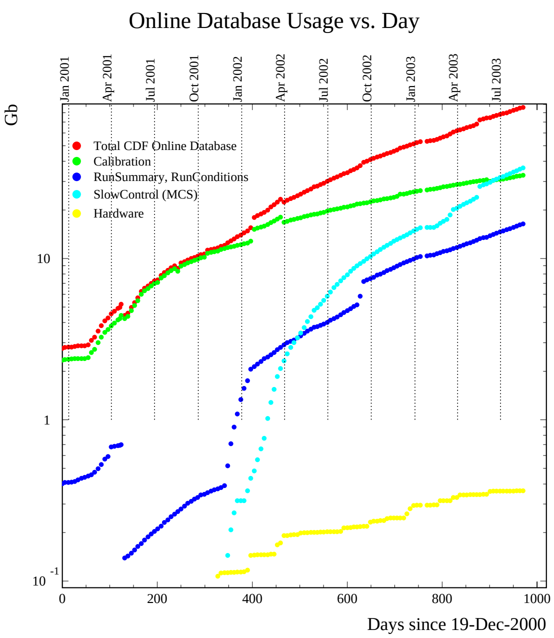
In FY2003 CDF purchased a replacement for fcdfora1. The new system is a Sun V880 with 8 900 MHz processors, 32 GB RAM and about 1 TB of fiber channel SCSI drives, as well as a giga-bit NIC. The replacement system has arrived on August 15 2003. The cost of the system is $64,460. An additional 1 TB of SCSI fiber channel RAID disks was also bought bringing the total disk space to 2 TB. The switch-over will occur during the September 2003 accelerator shutdown. The old fcdfora1 will then be deployed as a stand-by database for the online system in order to increase robustness.

Two Dell 2650 servers with dual Xeon 2.8 GHz PIV processors with 2 GB RAM and 72 GB of disk each were purchased to host CDF DB Browser.

[1] Sequential Access through Metadata
[2] *used* and *allocated* space for the tables

name	OS	CPU	RAM	Disk	Oracle version
b0dau35	Solaris 2.7	2×400 MHz USparc	1 GB	2.4 TB	9.2.0.3
b0dau36	Solaris 2.8	2×400 MHz USparc	4 GB	1.6 TB	9.2.0.3
fcdfora1(2)	Solaris 2.8	2×400 MHz USparc	1.25 GB	500 GB	8.1.7.4
fcdflnx1	RH AS	4×700 MHz PIII Xeon	4 GB	1 TB	9.2.0.3
fcdfora3	SuSE ES 8	2×2.8 GHz PIV Xeon	6 GB	1 TB	9.2.0.3

Table 20: Database hardware and software configuration



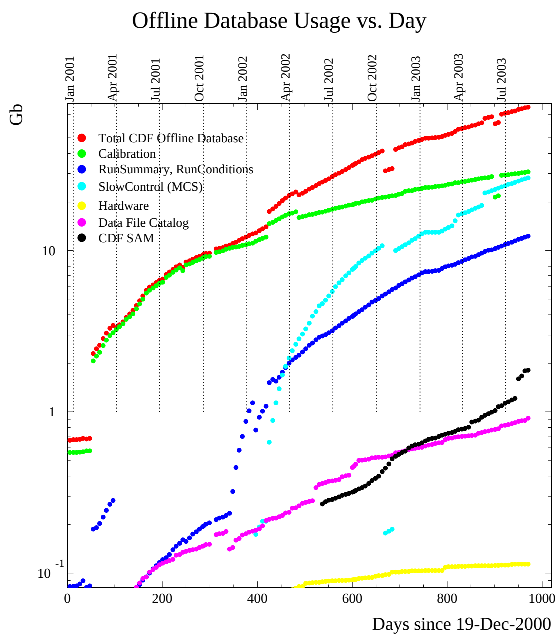


Figure 12: CDF DB space usage on online and offline production Oracle instances

7.2 Database Replication Hardware

The online data logger and production farms need continuous access to the offline production database in order to log and reconstruct raw data. At the same time, until September 2002, the offline production database was also accessed by users running analysis jobs primarily in read only mode. Increased analysis activity accompanied by substantial growth of CPU power led to several incidents when database and system resources could not handle the demand. This issue was addressed by developing a strategy of distributed databases via replication on site and at remote institutions. The replica copies of the database are read only instances accessed by the majority of users while the production farms and online data logger have exclusive access to the primary offline production database. As pure read-only databases the backup costs are minimal.

The first replica, cdfrep01, hosted by fcdflnx1, running Red Hat Advanced Server Linux operating system, was implemented in the summer 2002. The replication proceeds via Oracle read-only replication with either online or offline production databases as sources depending on the application. In addition, replication development work using Oracle streams replication is performed using on of the first generation CAF fileservers, fcdfdata012. Cdfrep01 is used by the CAF and all other users. The offline production machine has become isolated from disturbance by general users and is exclusively accessed by online data logger and production farms. There is a fail-over to fcdfora1 in case of emergency or maintenance work on fcfdlnx1.

In FY2003 CDF database group acquired a new replica host, fcdfora3, a dual PIV 2.8 GHz 6 GB RAM with about 1 TB of disk space on IDE RAID arrays, assembled from Maxtor-160 GB disks, attached to 3Ware IDE-RAID controllers. This host runs newer SuSE Enterprise Server 8 OS and hosts Oracle server version 9.2.0.3. The latter allows to use new Oracle replication mechanism – streams replication. The new instance name is cdfrep02.

Remote sites may choose to implement their own replicas, Oracle or freeware based, that contact cdfrep01 only, with no direct access to the primary offline and online databases by using Oracle streams replication or replication on demand using procedures developed by joint OSS-DS/CDF project. Expressions of interest have been made by TTU, Glasgow, Alberta, Karlsruhe and Tsukuba, with more to follow. Note that the remote institutions are expected to provide their own hardware and at least one responsible person, although assistance will be available from Fermilab.

FY	DB CPU (n-ways)	DB Disk (TB)	Cost ($M)
2004	2	6	0.10
2005	2	9	0.10
2006	2	14	0.10

Table 21: Database CPU and disk procurement plan. The fiscal year, the number of n-way Linux boxes purchased that year for DB machines, the TB of disk purchased and the cost.

The current configuration is that Fcdflnx1 is a primary replica machine and fcdfora3 is about to become a secondary replica set up with load-sharing and fail-over to fcdflnx1. In FY03 we set up these two DB replicas primarily to support the additional connections from the CAF. The plans for future purchases for db machines are shown in table 21. In FY04 we plan to purchase a replacement to the aging fcdflnx1 and one additional replica machine, both to support the increased demands from an expanding CAF. Similarly in FY05 and FY06 we envision that adding two replicas each fiscal year, or upgrading the existing replicas with more powerful machines, should keep up with the expanding number of total connections from the CAF, the remote institutions and the trailers.

7.3 Database Replication software

The current plan for replicas on-site and remote is to use Oracle replication wrapped with some convenience scripts and utilities. For the initial replica, cdfrep01, we are using Oracle read-only replication. Read-Only replication cannot be extended to remote nor trailer sites because of the online firewall. Read-Only replication also does not scale well with many replicas as the replication can be setup only from the originating database, that is, it is impossible to setup sequential replication or replication of replica. Any DDL change in replicated database will have to be manually propagated to many replicas resulting in prohibitive administrative overhead.

Replication capabilities to remote and non-FCC sites can be extended by deploying Oracle streams functionality which is available starting from release 9.2.0. This release has arrived in summer 2002. Oracle streams will allow the data propagation to proceed in sequential mode, thus avoiding firewall issues because the source can be the offline rather than the online database. In addition, the sequential nature of replication imposes a much lighter load on the source of replication. Streams also allows automatic propagation of DDL changes to replica sites.

Streams replication would thus make it practical to consider maintaining offline database copies based on Oracle at remote institutions, for example in Japan, the U.K. and Italy, for which latency and speed of access to the databases can be limiting factors.

For about a year the CDF data base group together with CSS-DS department of Computing Division have been working on testing the features of Oracle streams replication and developing tools that would facilitate setting up and running of such replication. Initial results are not very encouraging as it seems that there is some performance penalty compared to read-only replication, and bugs in the Oracle procedures. Work is ongoing, and the Oracle team is very responsive leading us to believe that maybe at least partial replication will be possible to set up using Oracle streams. That is, only the tables necessary for an analysis job would be replicated.

We have completed a project that explored the possibility of setting up and running of freeware database as one of the solutions specifically tailored at remote institutions. The subset of tables necessary to run typical analysis job has been identified; the set of scripts and procedures to select data from these tables, convert Oracle specific data types into ASCII, setup freeware[3] DB schemas has been created. CDF DB access API has been furnished with mySQL back-end. There exists two freeware replicas of offline production database - mySQL and PostgreSQL. Physicists at Karlsruhe University (Germany) were able to run jobs on high statistics data samples accessing information at Fermilab mySQL server. Several issues have been identified and being worked on.

To complement freeware development and simplify code maintenance there is an ongoing effort to implement single ODBC back-end to CDF DB access API. Once completed, this will allow us to use any ODBC compliant database implementation as a replica instance at remote institutions or at Fermilab.

CDF has expressed interest in the n-tier database access currently used and developed by D0 collaboration and Fermilab CD. Introduction of the n-tier access resolves Oracle licensing issue and provides a local caching or "secondary sourcing" of data from offline production database resulting in more efficient use of computing resources at the remote sites. CDF database group participated in formulation of requirements to the design of n-tier DB access system for CDF.

7.4 Interaction with Online System Improvements

During the past year, the CDF offline system experienced one major interruption lasting nearly 2 days to accessibility to the calibration database. This interruption was principally due to a deletion of the online database tablespaces for a large fraction of this application schema that cascaded to the offline system because of the explicit parent-child relationship between the online and offline copies.

We are studying the degree to which the use of Data Streams replication as described in the sections above can improve the robustness of the replication scheme against such possible failures. This appears to be a large potential advantage of Streams replication. In addition, the online system managers with help from the Computing Division Database Systems Group and CDF database group are studying the possibility of implementing advanced methods

[3]mySQL and PostgreSQL DB implementations were exercised

for improving online system time-to-recovery and reliability through systems such as Oracle Data Guard, which maintains a hot standby of the primary database. The fcdfora1 machine can be used as such stand-by dataset. The testing can be performed using existing fcdfsun1 machine currently running SAM DB server, which is going to be moved to the Linux box.

If the methods mentioned above work out, we may be in a position to study their application to protection of the read-write capability of the SAM and DFC databases in the offline system. The hardware and hot-standby instance for this step would need to be separate from (i.e., in addition to) that for the online system.

8 Networking

8.1 FCC LAN

The CDF LAN grew significantly in FY02 and FY03 with the introduction of a network based computing model: network attached tape drives using Enstore and network accessible disk and CPU via the CAF. We anticipate similar growth in the future. Figure 13 shows a simplified and incomplete diagram of the current and anticipated CDF LAN with major components.

The heart of the current CDF network is the offline switch, a CISCO 6509 in FCC2. Fcdfsgi2 is currently connected to this switch via 5 Gbit connections, 1 for interactive use and 4 for Enstore. The CDFEN Enstore robot currently has 10 Fast Ethernet (FE) connections to the offline switch for the movers for T9940A drives, and 10 GigE connections to the offline switch for the movers for T9940B drives. The stage 1 CAF fileservers use 16 GBit connections and the CAF stage 1 worker nodes use 67 FE connections to the offline switch. The stage 2 CAF has its own 6513 switch connected to 76 CAF fileservers via GBit connections and connected to 217 worker nodes via FE connections. The farms write to Enstore via a dual GB/s connection between the offline switch and the farms switch, another CISCO 6509. The farms worker nodes are each connected to the farms switch via FE, except for the worker nodes that will serve as input and output nodes, currently 8 input and 8 output, which have Gbit connections. The raw data logging node fcdfsgi1 directly mounts the dual ported disk with b0dau32 via Gbit connections. The offline switch is currently connected to the FNAL switch, and then to the outside world, via a single Gbit connection. The offline switch is also currently connected to the trailers switch, a CISCO 6509, via two Gbit connections. The desktops in the trailers are connected to the trailers switch using FE connections, although sometimes indirectly via satellite switches.

The needs of the CAF and disk cache have been driving the LAN plans shown in table 22. For stage 1 in FY2002 we purchased two 48-port FE modules and one 16-port GigE module and attached them to the offline switch. We also upgraded the switch capacity (fabric) to 256 GBit/s. For stage 2, later in FY02, we purchased an additional five 48-port FE modules and five 16-port GigE modules and attached them to a new CISCO 6513 switch.

For FY03 we have initiated the purchase of another CISCO 6513 switch for the CAF stage 3, along with seven 48-port 10/100/1000 GigE modules to support 231 worker nodes and 19 fileservers (2 ports each) that are on order. The connection between the offline switch and the CAF stage 2 switch is being upgraded to 10 Gb/s from 3 Gb/s in FY2003, and the CAF stage 3 switch will start out with a 10 Gb/s connection to the offline switch. The offline switch is also being upgraded to support the new 10/100/1000 copper GigE technology. The FY03 expenditure of $220K for networking resulted from $40K of additional gigE and FE

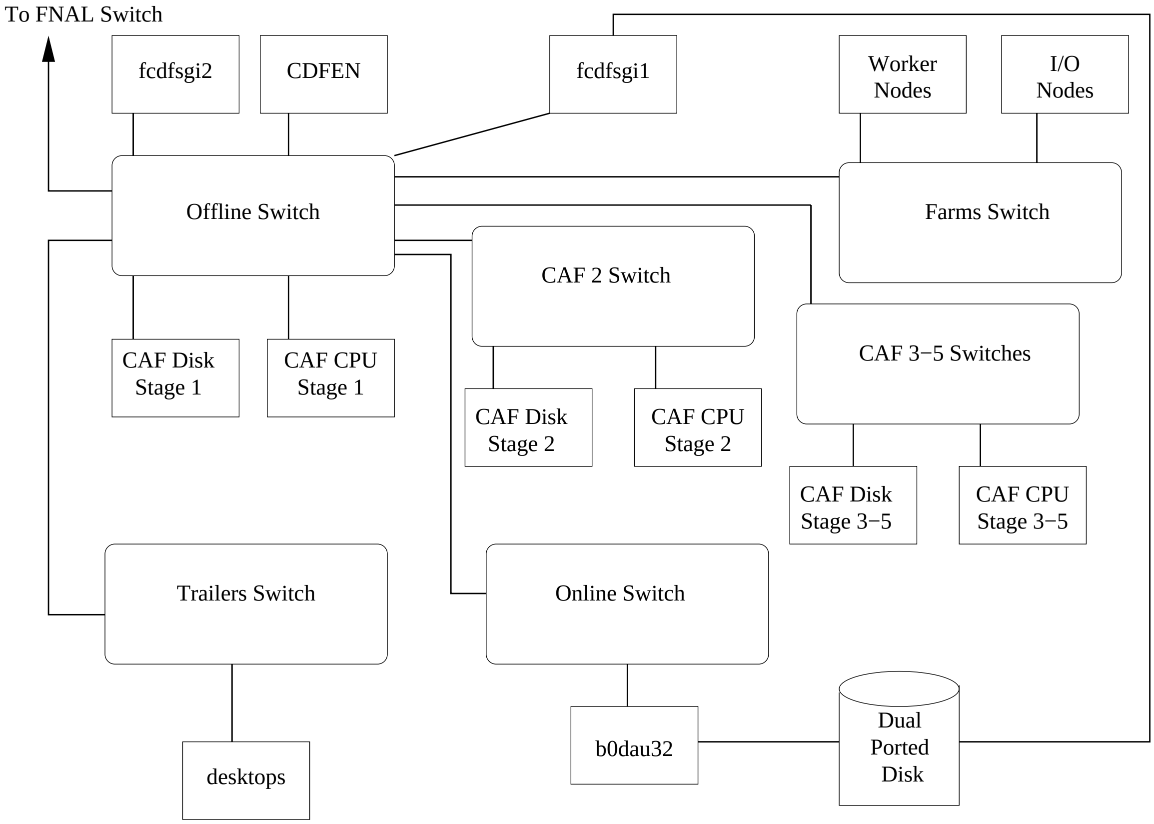


Figure 13: A simplified view of the CDF Local Area Network. Major computing elements in boxes, switches in rounded boxes, and Ethernet, GBit, or multi-GBit connections are shown by lines.

connections for the offline switch, $110K for the stage 3 CAF switch and its gigE and 10 Gb/s modules, and $70K to upgrade both the offline switch and the stage 2 CAF switch to use copper gigE and to supply them with 10 Gb/s modules for inter-switch connectivity. All in FCC.

For each of FY04 through FY06 we plan on purchasing additional CAF switches with the necessary modules for connecting worker nodes and fileservers. In table 22 we estimate the cost of this networking by assuming a Moore's law like decrease: networking costs that drop by a factor of 2 every 18 months. Note that this may be hard to achieve in practice, the drop from FY02 to FY03 being significantly less than Moore's law due to the cost of upgrading all the computer center networking infrastructure to use new copper gigE and 10 Gb/s technology. Implicit in the networking costs are an assumption that the bulk of the computing center upgrades are complete. Networking costs for the reconstruction farms are discussed in section 6 and included in the farms total costs. The other major cost anticipate for the LAN is the trailers, discussed in the next section.

8.2 Trailer LAN

The networking in the trailers has not been upgraded in many years, and the networking group has recommended an upgrade for each of the past 3 years. The CDF trailers LAN currently supports 100 Mb/s connections to the majority of CDF offices, and this lags behind the current network capabilities of desktop Ethernet cards which are 1 Gb/s, and restricts

FY	FCC Cost ($M)	Trailer Cost ($M)	Total Cost ($M)
02	0.25	0.00	0.25
03	0.23	0.00	0.23
04	0.14	0.11	0.25
05	0.09	0.10	0.19
06	0.06	0.06	0.12

Table 22: LAN procurement plan. The fiscal year, cost of Fermilab computing center networking, cost of CDF trailers networking and total cost.

the data transfer rates for existing fileservers in the trailers. Currently multiple satellite switches are used to extend the ports available on the trailers 6509 switch, in an architecture that lowers the bandwidth capacity of many offices. Since there is essentially no more room on the existing 6509 switch for additional ports, more switches would need to be added to support further expansion.

There are currently two Gb/s connections between the trailers LAN and FCC. The typical usage of this link is 200 Mb/s with frequent spikes in usage to around 600 Mb/s. The usage of the link between the trailers and FCC has been increasing steadily during the last year. People use their desktops for analysis and access datasets in FCC via dCache or rootd. The link was doubled from 1 Gb/s to 2 Gb/s in FY03. We should anticipate the demand for bandwidth over this link to scale with dataset size, and do not expect that trunking multiple Gigabit lines (limited to 4) will work for more than one more fiscal year. We must have networking in the trailers that will support 10 Gb/s connections between the trailers and FCC.

For all the aforementioned reasons we believe now is a good time to begin an upgrade of the networking in the CDF trailers. An appropriate upgrade to the trailers networking, giving it sufficient connectivity to the computing center for the next 3 years, allowing it to support more connections, and allowing it to support Gigabit connections to each office would cost roughly $420K if it were done in FY04. We plan to do it in stages over FY04-FY06 for a total cost of roughly $270K.

Here is an example of one way to do the upgrade, although others should be carefully considered, and cost estimates if it were all done in FY04. We need to buy a 6513 switch with supervisor module for around $34K to both enable the upgrade to gradually take place and to provide more connections. There are 168 offices with fiber connections to the trailer LAN, 64 offices with copper connections to the LAN (roughly 2 connections per office), and there are another 168 connections in the new office building. The 168 fiber offices would each need a $800 1 Gb/s desktop switch to drive Gb connections: total cost of $134K. The 168 fiber offices would require 7 24-port Gigabit modules at a cost of $10K per module, or $70K total. The 64 copper offices could use the original 6509 switch, which would need a new supervisor module for $20K. The 64 cooper offices probably host around 128 connections, which will require three 48-port 10/100/100 modules for a total of $32K. The new building will need its own 6509 switch for $30K and $42K worth of 10/100/1000 modules to support 168 copper Gigabit connections. Inter-switch connectivity would require four $14K 10 Gb/s module, one for each trailer switch and one more for the FCC switch, at a total cost of $56K. Since the 10 MB/s modules have 4-ports we would have 2-port to connect the trailers and

FCC which would allow for up to 20 Gb/s: ten times the maximum today.

In FY04 we plan to buy the 6513 switch and supervisor module that networking has been recommending as a foundation to upgrade the trailers networking, and also upgrade the existing 6509 to handle 10/100/1000 technology. We also want to buy at least one 24-port GigE module to allow some fiber offices to have Gigabit connections, and at least one 10/100/1000 module to allow some copper offices to have Gigabit, and three 10 Gb/s modules to increase the connectivity between the three switches (trailer 6509, trailer 6513, fcc 6509) to 10 Gb/s. This is a total cost of $110K in FY04. The remaining $310K cost of the upgrade we will split into equal halves for FY05 and FY06 and apply a Moore's law reduction in cost to obtain the costs for FY05 and FY06 to complete the trailer upgrade.

8.3 WAN

The traffic on the up-link between the CDF offline switch and the main FCC switch is shown in Fig. 14. This sets the scale for the amount of bandwidth that could be going offsite from CDF.

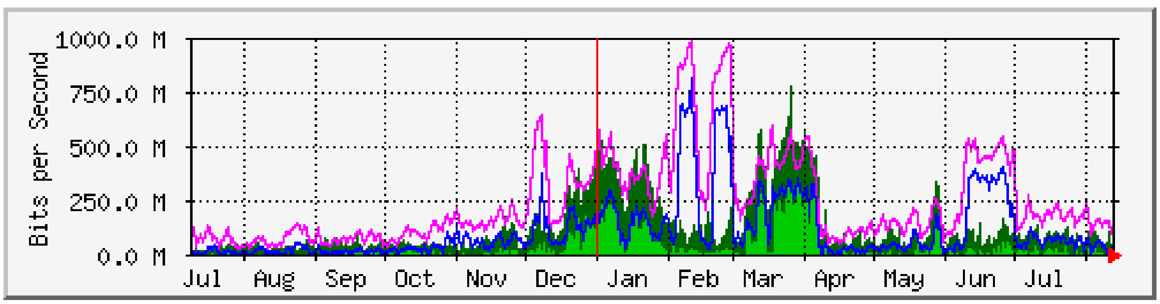


Figure 14: The traffic for the last year on the Gb/s up-link between the CDF offline switch and the main FCC switch.

In FY03 the OC3 connection between Fermilab and ESNET was upgraded to OC12 with a capacity of 622 Mb/s. In FY04 we are expecting the lab to request an OC48 with a capacity of 2.5 Gb/s, and around FY06 we are anticipating a request of 2 OC192 or 20 Gb/s. As recommended by the previous review [7], this connectivity is being augmented with a direct link from Fermilab to Starlight that should provide additional capacity beginning around March 2004. Should we saturate the ESNET OC12, the starlight link should provide additional capacity of 4 Gb/s or more, which we hope will be routed through ESNET at starlight. There is a purchase order in place now to lease the use of the fibers from here to starlight, and depending on the optoelectronics the capacity of the link could be five times higher than the nominal 4 Gb/s expected. We are optimistic about the WAN capability from FCC to offsite in the future. Estimates from the previous review [1] indicate that the WAN needs of CDF are around 1 Gb/s, and these should be easily met quite soon.

9 Computing Facilities Outside Fermilab

9.1 History and Perspective

Since Run 1, CDF collaborating institutions have a history of using significant computing resources outside those provided by FNAL, either located at home sites or in the Fermilab trailers. This trend has continued into Run 2, and it had for many years escaped any programming and accounting, mostly because, as long as those resources have been exclusively used by single collaborating institutions, there was no way to track down their amount and effectiveness. Effectiveness of off site computing facilities had been limited in the past also by lack of automated data transfer (hand carried tapes only) and of tools for cross-site user's authentication and remote job submission. But this has started to change already and there has been a significant shift toward the deployment of significant off-site computing resources for MC production and data analysis dedicated to the CDF collaboration, following the last Fermilab Computing Review (June 2002). This has come about largely through the availability of large Beowulf clusters in several countries (Canada and the UK) with significant capacity dedicated to CDF, and the improved WAN bandwidths with consistently $>$ 100 Mbps throughput between the off-site centers and Fermilab. These has been possible because of the following several reasons. In no particular order:

- WAN speed has enormously increased in the last years and practically no data has been moved on tape to/from Fermilab during Run 2
- we have much friendlier and robust means to distribute CDF code that make it possible to insure reproducibility of results from the FNAL main platform to any remote laptop
- tools (like SAM) have become available to allow transparent access to remote data
- tools (like CDF CAF) have allowed already to build several computing farms that provide uniform access from any CDF user's location to any of those farms
- tools (like JIM) are becoming available to allow transparent automated brokering of user's job among many farms
- implementing and operating transparent access to FNAL based data and code repositories and computing resources for remote users, while it is working enormously better then in Run 1, is proving to be expensive enough human and money wise that it starts making sense to ask remote institution to repay this cost at least by allowing symmetrical access to their local computers
- driven by the political and technological developments in preparation for LHC era, most collaborators start to have access to large computer facilities that are built around the world and which may both have spare capacity in the near future and provide an easy path for the addition of CDF-owned resources
- several CDF institutions are already starting to open access to their local resources to all the collaboration, up to having formal MOU's for that, something that never happened in Run 1

Putting all the above together, both in terms of tools and of contribution/retribution, is the prerequisite for developing a carefully crafted political arrangement among the many CDF collaborators, which have widely different patterns of financing, contribution to CDF in other forms, and history in the collaboration.

We believe it will be more effective for CDF to build its "GRID" infrastructure using a bottom-up approach, were we start with our working system at Fermilab and add in remote resources a bit at a time, after having developed and demonstrated the effectiveness of the tools for the users, rather then to try to define a framework of "contributions" before knowing whether and how much those remote resources can be used. As we try to add value to our working environment, it is important to remember that any new tool or feature must give to the end user (the "famous thesis-hungry student") a significant improvement in terms of "time-to-publication", or no-one will use it.

Moreover the spread of new powerful computer centers and the affordability of computer farm components, is making possible for many CDF institutions to have private computing resources quite larger then they used to have in Run 1. Analysis and simulation work is moving away from FNAL to locally owned machines. This is also a welcome contribution as it reduces load on Fermilab systems, even in lack of a Grid environment. It is very difficult to quantify these things into "savings" or "contributions" in dollars. Again a flexible incremental approach based on reality as we go on will be the best way to manage this process and incorporate the resources into a coherent scheme.

In practical terms, we expect CDF to extend from Fermilab to incorporating computing facilities outside the laboratory in a step-by-step process:

1. allow easy usage of remote facilities for code development, data analysis and MC generation by single institution
2. move off site (part of) organized MC production
3. move off site (part of) single user MC production spreading it uniformly across all institutions
4. move off site analysis of secondary and/or tertiary data sets, by duplicating CDF data at remote institutions and giving everybody access to it
5. exploit large off site CPU capability for interactive analysis and reprocessing
6. develop formal agreement and define the "price" of each service, only after it has been demonstrated to work

Step 1 is done, and currently in use, although transparent access to large CDF data samples via SAM is still in development/test phase. Step 2 has started. Organized Monte Carlo for Summer Conferences 2003 was generated entirely outside of FNAL, as discussed in detail in Section 9.3. Step 3 is the near term focus and we expect this to start happening by Spring of 2004. Step 4 well be a combination of success with SAM (in 1.) and 3. Step 5 Is at present an R&D project. We expect that Step 6 will not lead to significant decreases in the FNAL computing budget as described in this report until FY06 given that the budget is driven by costs associated with Step 4 and Step 5 type of services, as well as tape archiving.

To arrive at a "Virtual Center" accounting model by FY06 without sacrificing service levels, significant human resource effort has to be expended in FY04 and FY05. FY04 needs to be focused on technical aspects of providing the services, including initial prototype deployments. FY05 is then focused on operational aspects, and thus understanding of "pricing" that reflects services rendered in Step 4 and 5.

9.2 Status of available off site resource

While some CDF collaborators own CDF-reserved computers, other share access to largish facilities with other experiments, this makes it almost impossible to tell a-priori e.g. how much CPU power CDF physicists can use off site. We expect that once a framework will have been clearly defined for usage of those facilities by all CDF collaborators, something like a minimum amount of CPU cycles available can be defined (Canada is doing this already, e.g.), while most likely accounting of off site facilities contribution will have to be done a-posteriori, based also on the actual efficiency and effectiveness of each single installation.

To give an idea of the amount of resources that CDF can tap into by accessing off site facilities, the following table, which has no pretense of being an exhaustive account, reports in no particular order the approximated amount of resources available by end of 2003 and a reasonable estimate for 2004 (most institutions still do not have an approved budget for then) in a representative set of institutions. Note that most sites are geared toward MC production especially as common CDF resources are concerned, so the local amount of disk is little informative, we write 0.1TB as storage in this case to mean that those sites do not envision allowing significant data access to CDF users at large, at least at present. Local access policies are still being defined and will possibly change in the future as a result of political negotiations and our attempt to enforce a common policy.

	2003		**2004**			
	CPU (GHz)	Disk (TB)	CPU (GHz)	Disk (TB)	Access Gbit/sec	notes
Canada	250	0.1	250	0.1	2	(1) (3) (4)
Canada	1000	0.1	1500	0.1	2	(2) (3) (4)
Germany (FZKA)	680	14	750	22	1	(2) (4)
Italy (INFN)	250	8	700	20	2	(1) (5)
Japan (3 sites)	116	13	120	15	1	(1) (5)
Korea	400	0.1	500	1	0.1	(2) (5)
Spain (ICFA)	130	0.1	200	0.1	0.6	(2)
UK (ScotGrid+RAL)	150	5	230	5	1	(2)
UK	200	24	200	24	1	(1) (5)
TexasTech	30	1	250	1	1	(2) (5)
Rutgers	54	7	110	15	0.2	(1) (5)
MIT	-	-	100	0.1	1	(1) (3) (4)
UCSD	-	-	100	0.1	1	(1) (3) (4)
TOTAL	> 3000	> 60	> 5000	> 90	-	

Table 23: Notes: (1) reserved for CDF group (2) shared with other experiments (3) dedicated to MC production (4) allows unrestricted equal access to all CDF members (5) leaves CPU cycles unused by owners to the rest of CDF for MC

It is clear that most sites are building local resources mostly for lots of CPU, rather then storage, according to the general perception that the most effective way to use off site resources will be user Monte Carlo production and reiterated analysis of smallish data sub-samples, while the skimming of large inclusive samples will best be dealt with at Fermilab.

Most off site institutions have a high speed local connection to the Internet, so access to those facility can be highly efficient, and Fermilab should make sure it is not going to be the bottleneck for that access even when data is moved to/from several off site locations at the same time. Experience shows that effective throughput on WAN can be limited by many hard to find bottlenecks. Our best experience so far is with the Canada-FNAL link that has shown reliably and consistently $\sim$ 200Mbit/sec for production MC transfer. which seem to represent the maximum presently allowed by Fermilab-ESNET connection.

9.3 MC Production

At present, the primary use of off site computing resources for CDF as a whole is in the form of off site MC production, mostly in Toronto and the UK. This is presently done by running MC jobs locally (i.e. no grid-like remote central control) and importing data back to FNAL by ftp to on site disks and hand write to Enstore. The Canadians have MOU responsibility for coordinating the MC production, as well as providing 1 Million events per day capacity in Canada; above the guaranteed minimum of 1 Million events/day, the Canadian cluster may be, and in past has been, able to produce significantly higher rate of MC production. Other institutions (UK e.g.) are considering the possibility to setup analogous MOU's for taking up responsibility to produce a fraction of the CDF needed MC events. The MC production group has already produced $>$ 100 Million events in FY03, about 60% in Canada and 30% in UK, and $\geq$ 250 Million events will be the likely total by the end of this calendar year. We expect this to increase more than linearly with luminosity as we improve the accuracy of the simulation, and physics analysis of Run 2 data matures.

All MC generated in this fashion is coordinated via the physics groups, but this is by no means all the MC generated by CDF. Most analyses need very large MC productions that are tuned to that particular topic and can not be shard with others. These "single-user" MC samples are, and will be in the future, produced in a chaotic way by singles or small groups as needed and will not be managed by the physics groups.

The next steps for general CDF computing off site is therefore providing users the means to generate their private MC samples, e.g. exploiting CAF installations at remote institutions. At present significant amounts of user level MC are produced on the FNAL CAF, which is a poor usage of a system built for good data access and tightly coupled to the main CDF data repository. Most of that work could be very well done off site and results copied back to FNAL. This will be easily accomplished using the current CAF tools, and we are working on a scheme to make this operationally reliable and reasonably painless for the site administrator. The hurdles for this are largely a matter of policy. As a ground breaker, we are presently working out MOU level agreements with MIT and UCSD to support these efforts using old computing equipment that was decommissioned out of Level3 and the production farm (this equipment is what is listed in the previous table as available CPU power at those sites in 2004). We further plan, as our grid project mature, to expand these sites into more general user analysis centers off site. Other sites, e.g. Rutgers, INFN, Japan and Germany, have expressed willingness to offer spare CPU cycles on their cluster to CDF in general and are expected to offer batch queues to all CDF members on an experimental way by the end of 2003, independently of possible formal agreements.

9.4 Incorporating off site resources: toward a CDF Grid

9.4.1 The Vision

The long term vision for CDF computing is that users develop and debug their application at their desktop somewhere in the world. They then submit their job to SAM-grid, specifying a dataset to analyze in addition to the usual CAF information. SAM-grid selects an execution site based on locally available data as well as CPU resources. The user job is queued at the local site, eventually instances of it start. SAM provides input data, and the user writes out their output into a local scratch area. After completion of an instance, the user may declare the files produced to the DH system for storage in a location where the user has sufficient quota to store the files on disk.

It is up to the DH system to track the files from the moment they are declared until they arrive at their "permanent" disk location. The latter is generally a site different from the execution site, and different users have their disk quotas generally implemented on different sites. All the data movement as well as final location is completely transparent to the user. In particular, the DH system is responsible for randomizing file location for files of the same user and output dataset across multiple physical devices in order to guarantee large I/O bandwidth, and minimize chances for congestion due to access patterns. This access optimization is generally performed by the storage layer rather than the "data handling" layer of the DH system.

Similarly, it is the responsibility of DH to optimize CPU utilization of off site clusters by prestaging data appropriately thus minimizing idle times due to data unavailability. In addition, DH provides metadata catalogue services for the user's output data such that it can be reused as input for a future job by anybody in CDF.

Data stored in this manner is not backed up to tape, and may be permanently erased by the user who owns the data. Tape archiving would require in general additional concatenation to achieve file sizes for efficient operations of tape archive resources. We presently allow users to archive their data on tape free of charge. This service is not widely used. We may change our policy with regard to tape costs if usage of this service were to increase significantly.

As organizational principle we expect to be guided by the notion of physics centers rather than regional centers. Assuming we can arrange policies of ownership for resources at the level of "virtual center" it is much more efficient to concentrate a given dataset and the CPU resources required for its analysis in a set of dedicated sites rather than spreading all datasets across all sites more or less evenly. Needless to say, this ideal will require some amount of negotiation and deliberation to be successful.

9.4.2 The Tools

Technically, we will accomplish this via a combination of SAM, dCache, Enstore, CAF, and JIM. At present, dCache read operations, Enstore, and CAF are fully in production. SAM is expected to follow in FY04. JIM will be deployed during FY04 with production level support in FY05. Write operations to fully virtualized disk space, as well as dCache write pools are expected to follow in FY04. In addition to this batch based processing, we are also pursuing interactive computing, both in form of a central interactive platform at FNAL discussed in section 4, as well as an interactive grid.

The interactive grid computing effort is a collaboration between UCSD, INFN, and MIT. The goal is to build an interactive grid computing system with response times to queries of order 10s for analyzing O(10GB) ntuples. The system is to be based on Root's PROOF tool by adding an interface to SAM metadata catalogue, and Condor/Globus grid middleware. A first prototype is expected for SC2003 on November 15th.

The driving force behind the interactive grid concept is to provide the CDF users a strong incentive to optimize their analysis for early transition to ntuples, performing only the initial processing steps within our AC++ production environment. The CPU and thus $$ savings of this approach, if successful could be substantial. Typical event rates for a user analysis based on ntuples exceeds that based on AC++ by 2-3 orders of magnitude.

9.4.3 The Financial Side

The CDF International Finance Committee is currently debating the first approach to the formalization of foreign contribution to CDF costs.

Up to today, even in lack of such a formalization, CDF has nevertheless received substantial financial contribution by foreign countries, Japan and Italy mainly, who have done so since many years, however more recently also UK, Switzerland, Korea and Canada have contributed. In particular for what concerns computing, Canada has taken on a serious commitment to MC production and is providing very significant computing resources for it, and so is the UK and many US institutions who have produced Monte Carlo samples for CDF data analysis since roughly December 2002. It is CDF's first priority to preserve all positive sides of how things have been working till now, and therefore to be very cautious and careful in defining a brand new written policy. For these reasons we do not expect the International Finance Committee to reach a final document on this topic before the end of FY03 and expect FY04 off site contributions to CDF to keep coming as it happened so far, in particular financial contribution to CDF computing will come in form of purchase of hardware to be installed in the FNAL CAF.

FY05 off site contributions may come in form of CPU/disk usage off site by the CDF collaboration as a whole. Such contributions will be accounted for as common fund contributions in a manner that will be defined by the CDF International Finance Committee.

Until our grid vision is fully implemented, and shown to be fully functional, we consider it prudent to require that baseline needs of CDF are fulfilled by on site computing at Fermilab. Assuming that the present developments are indeed successful it is conceivable that we start a more global computing cost accounting in FY06 or FY07.

9.5 Offsite Productivity Issues

A common theme among University based institutions in CDF is a concern about being able to attract and retain the best young graduate students during their first year when they cannot be at the lab. This has been identified as one of the most important challenges in our field by several of the US collaborators. The CDF software is perceived to be sufficiently complex, and frustrating that it makes it very hard to motivate young students. In many cases the students leave the field of HEP to work on something that has quicker entry into physics and is thus more attractive.

There is a clear consensus that a focus on user productivity is required for the upcoming years. It is however less clear who should bear the burdon of providing the human resources required to make substantial progress. A compromise might include increased FNAL computing division support for software tools like debugger, profiler, memory leak checker, and code browser, in combination with collaborator support possibly via offline shifters for improved documentation, and an effort on making programming interfaces more consistent, as well as defining default analysis objects like tracks, photons, leptons, etc. which are well understood and documented.

A related set of concerns revolve around making the production output generally useable for the enduser, with well documented, and understood quality of the data.

10 Offline Analysis Introduction

The most significant short term activity in the Offline Analysis is preparation of a new release of the CDF software, and a complete rereconstruction of all the CDF data with this new release between September-December 2003.

This release is not simply a major effort but also marks a turning point, in that for the first time the quality of the reconstruction and simulation is expected to exceed Run 1 in almost all aspects.

Reconstruction software for sub-detectors that are new in Run 2, e.g. L00 and ToF, is generally less advanced. The ToF reconstruction will require significant effort during FY04 in order to become useful for flavor tagging. The L00 reconstruction is progressing and it may be ready for first use in physics analysis at the end of this year. The forward tracking using the silicon detector needs additional work in order to improve efficiency and purity before it will be useful for physics tasks such as b-tagging, though it already is useful for the tracking of isolated particles. Other examples include forward Muons that may become useful for physics for the first time in FY04. Future reprocessing of the entire dataset may be driven by improvements in the code for these sub-detectors, or by refinements in calibrations such as COT or silicon alignment.

In addition, we will use a compressed data format for the first time, and increase the level of dataset splitting out of production so as to avoid physics-group-based secondary dataset production based on Level-3 bits entirely. This will shorten the latency for creation of some of our largest secondary datasets (e.g. 2-track trigger for B group, jet triggers for QCD group) substantially.

Accordingly, rereprocessing, and data validation are special focus areas in the short term. Data validation both online and offline was recently taken on as an MOU responsibility by Barcelona. We expect this to provide the necessary additional resources to overcome present shortcomings in this area.

A second area of concern is user analysis support. The issues here range from ntuple or μ-DST support, to software tools support (debugger, profiler, memory leak checker), to general software documentation and ease of use. There is a general concensus in the collaboration, especially among offsite collaborators, that CDF has still a long way to go in this area before the full human resource potential of our 767 collaborators can be exploited. While we hope to make progress in this area within FY04, we anticipate a multi-year effort is required before this is no longer an issue for a substantial part of the collaboration.

For the upcoming year we expect to concentrate our efforts on ntuple and software tool support issues. We expect to accomplish the latter via "technology transfer" between the Applied Physics Software group in CEPA department of the FNAL-CD and the CDF collaboration. An ideal model for this might be to identify small projects like profiling and speeding up the CDF simulation executable, and assigning a CDF collaborator to a CD expert. The hope would be to create local experts in CDF who could then spread the expertise further.

To tackle the ntuple/μ-DST problem we intend to charge a task force inside CDF with evaluating options and arriving at a recommendation. Details remain to be worked out. The objective is to move more users earlier in their analysis away from the costly AC++ framework (5Hz event processing rate per GHz processor) to something that is 2-3 orders of magnitudes faster. Part of the charge to the task force is going to be to evaluate the impact on data handling, and to elucidate relative advantages of the different ntuple strategies

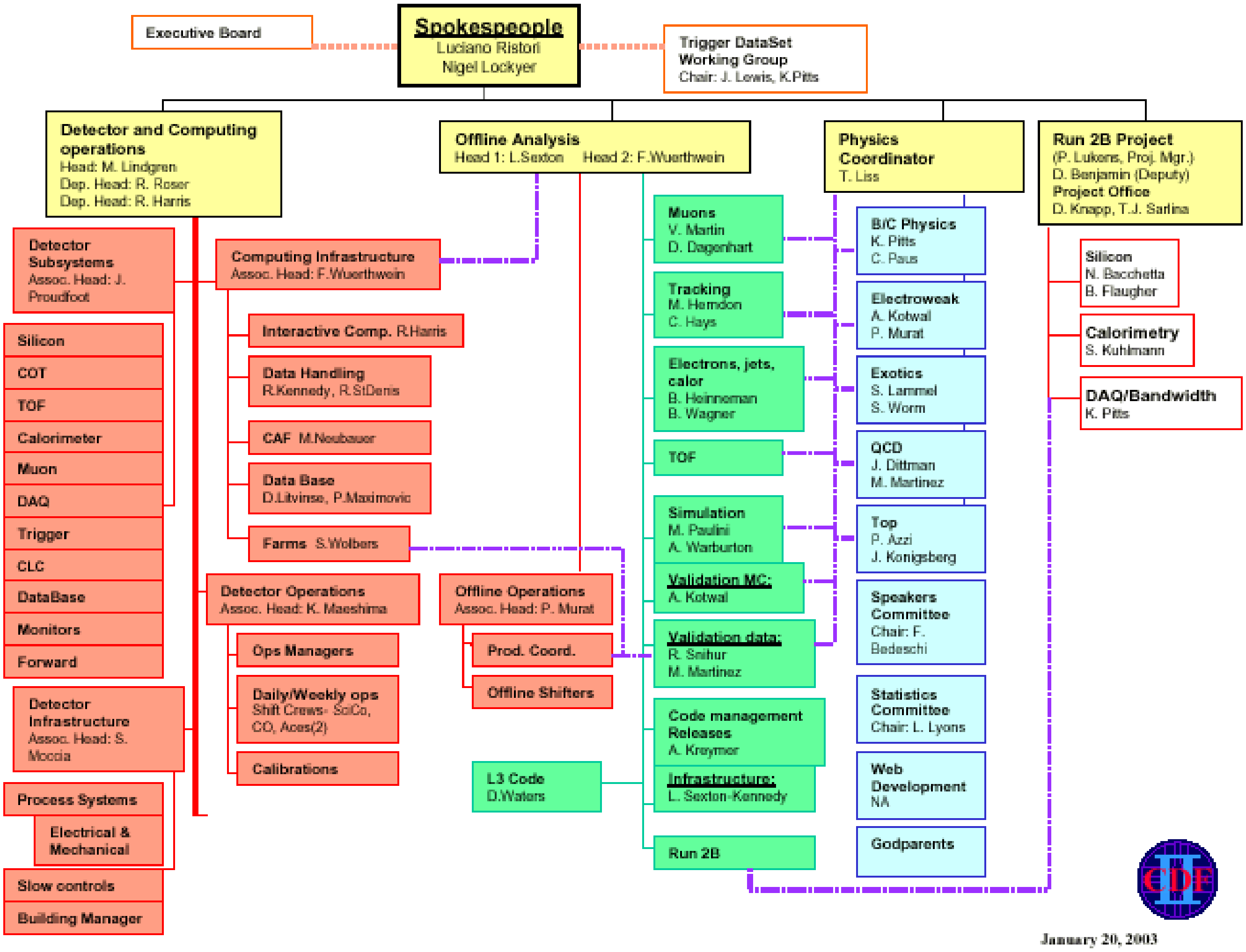


Figure 15: CDF organization chart as of July 2003.

presently in use at CDF.

10.1 CDF Computing Organization

Figure 15 shows the CDF organization chart. CDF computing consists of two parts in this chart, the computing infrastructure discussed previously in this report, and the "Offline Analysis" part which is about to follow. All equipment funds, and the bulk of the computing division support is expended as part of computing infrastructure, offline operations, code management, and software infrastructure. In addition to the computing division support, DH, CAF, DB, and Farms all receive substantial support from CDF collaborating institutions, which is essential to the success of these projects. Offline Analysis depends to a large extent on the workforce from CDF collaborating institutions.

To guarantee successful operations of CDF computing as a whole we need to guarantee very close horizontal relationships between projects in different arms of the organization, as indicated by the blue dot-dashed lines.

An obvious example of such a relationship is coorperation between Offline Operations, Data Validation, and the Production Farm to guarantee expedient processing of data of well understood quality.

Similarily, the physics groups of the experiment are of course an important stakeholder in the reconstruction and simulation software, as well as data quality control.

A maybe less obvious relationship that has proven crucial largely through its absence over the last year is the connection between Offline Analysis and Computing Infrastructure management. The latter controls all the computing resources, while the former controls the applications and products that consume the resources, and has a substantial need for resources, especially during major software release development cycles. Having a tight coupling here has proven very useful over the last couple months.

11 Tracking

The current version of the tracking reconstruction software performs with high efficiency and purity within reasonable time constraints. Reconstruction efficiency exceeds 99% in the COT for isolated tracks, and the efficiency for subsequently attaching silicon hits is above 90%, where most of the inefficiency is due to hardware failure. The outside-in tracking purity, i.e. the fraction of tracks with COT and silicon hits that have well-measured parameters, has been measured to be approximately 99% for isolated tracks. Recent improvements in the track reconstruction code include better alignment, raw data compression, recoding of the silicon tracking algorithms (resulting in 2% higher efficiency), and the implementation of new tracking algorithms, including inside-out tracking (from the silicon to the COT), calorimeter-seeded tracking, and the development of a specialized version of the silicon standalone tracking for use in triggering at Level 3.

Further advances are anticipated in the next year. The silicon stand-alone tracking has inefficiencies that can be reduced. The inner silicon layer (00) is in the commissioning stage, with the potential for first physics use this winter. Understanding of tracking errors and the track fit can be improved. In addition, the event reconstruction speed is dominated by the tracking code, with 2/3 of the total reconstruction time spent unpacking the raw silicon and COT data and reconstructing tracks. Approximately 1/3 of the event size is taken by the tracking detector raw data and the reconstructed tracks. Optimization has already taken place to reduce the tracking reconstruction time and the size of the raw detector data. With significant manpower, further improvements could be made with the most likely benefit from compressing the reconstructed tracks.

Aside from the primary goals of improving silicon standalone tracking, incorporating layer 00 and improving reconstruction speed and track collection size, a long-term project is optimizing reconstruction for a high luminosity environment.

The efforts in these areas come primarily from a number of university groups. Fermilab personnel have contributed recently by developing the compression and calorimeter-seeded tracking. Further development of the tracking code could use a significant increase in manpower, especially from the Fermilab group, which has traditionally contributed a higher proportion of manpower than is currently involved.

12 Muons

The muon reconstruction software for the main parts of the central muon subdetectors: the CMU, the CMP and the CMX, is fully functional. For high pT physics, the central muon chambers systems, the CMU and the CMP, which are situated at $|\eta| < 0.7$, are usually used in combination, *i.e.* a muon is required to leave a stub in both the CMU and CMP chambers. The measured value of the reconstruction efficiency for these muons is around 95%. The CMX chambers, situated at $0.7 < |\eta| < 1$, have a measured reconstruction efficiency of around 99%. A fourth muon chamber system, the BMU, is situated at $1 < |\eta| < 1.5$. Initial studies indicate that the efficiency of this system is around 90%, however some effects seen in the reconstruction are not fully understood.

There are two main priorities for the muon software over the next year. The first is to understand and improve the reconstruction of the other parts of the muon systems: CMU Wedge 17, the bluebeam section of the CMP, the CMX miniskirt sections and, most importantly, the BMU. This will involve tuning values for calibration constants, aligning the subdetectors in the reconstruction and validating the improved algorithms. One major improvement for the BMU reconstruction is to allow tracks with only a few silicon hits to be linked to BMU stubs (including silicon-only tracks). Another improvement to both the miniskirt and BMU reconstruction will be a modification of the BMU geometry to avoid overlapping geometrical containment volumes. The second task is to improve the simulation of the central muon detectors, such that the simulation reflects more accurately real data. This will involve using more accurate drift models, better modeling of the reconstruction effects along the chambers and taking account of dead channels.

There is also a number of smaller tasks that need to be completed: an upgrade of the muon software validation module, maintenance of the documentation, checking the detector alignment and integrating muon scintillator information and χ^2 quality variables into the reconstructed muon objects.

13 Electrons, Jets, Calorimetry

The calorimeter system in CDF has several components: the main calorimeters are used for the energy measurement of electromagnetically and hadronically interacting particles up to pseudo-rapidities of 5.5. Additionally the electromagnetic calorimeter is instrumented with wire chambers which aid the position determination and photon/electron identification. There is also a pre-shower detector which helps photon-pion separation.

The reconstruction software for all these detectors is essentially finished including calibration mechanisms using the CDF database which allow for time-dependent variations of the detector response. This includes not only the reconstruction of the bare quantities measured in the detector but also higher level particles: jets, electrons, photons. The reconstruction software is robust and is fast enough to not cause any delays for the end-user. Recently a lossless compressed format for all the calorimeter data was developed. In the past year the calorimeter reconstruction was frequently run by the users in order to get access to the latest calibrations. This proved to be feasible on the central analysis facilities of CDF. Compared to e.g. the banks and the reconstruction related to tracking the size and speed of calorimeter related objects and software are negligible.

14 Simulation

The CDF detector simulation is integrated into the AC++ application framework used to process events at CDF. The tracking of particles through matter is performed by the GEANT3 [8] package. The CDF software uses the same geometry for event reconstruction and simulation. The CDF geometry package provides the volume description and geometry tree creation. The CDF geometry was designed to allow seamless transition to GEANT4 in the future. An output event of the simulation contains data in the same format as raw detector data plus Monte Carlo truth information.

The main simulation executable, cdfSim, allows generation of physics events with different generators such as Herwig or PYTHIA. In addition to decay routines internal to each MC generator, particle decays can be simulated by customized decay packages such as QQ, EvtGen and Tauola. The cdfSim executable may be used with different configurations of subdetectors, different geometry levels, and physics processes depending on desired accuracy versus time efficiency. All sub-detectors of the real CDF detector can be simulated. In addition, the CDF geometry contains a detailed description of passive material elements, in particular within the silicon detector.

The current cdfSim executable represents a tunable framework that significantly improves the CDF detector simulation compared to Run I. For example, the simulation of the CDF silicon detectors allows to misalign the individual silicon wafers including surrounding support structures to allow Geant to trace through the same geometry as in the real CDF detector. After the simulation step, MC data are then reconstructed like real data using alignment tables to correct for residual misalignment. In the past couple of years, the main activity of the simulation group consisted in tuning the simulation response with Run II data. This has reached a level fully adequate for deriving physics results for conferences and for the first CDF Run II publications. Tuning of the detector response will continue with more sophistication as higher statistics Run II data become available. In particular, a complete re-tuning of the calorimeter response is currently ongoing and the description of the Silicon passive material requires more details.

The work of improving the CDF detector simulation has been traditionally carried out by university physicists such as students and post-docs in collaboration with the CDF physics groups. For example, the leadership of the simulation group in the past few years has been provided by Berkeley and Carnegie Mellon University(CMU). As another example, CMU provided one FTE for technical involvement in the simulation framework and maintenance of the cdfSim executable. Some support for particularly technical projects had occasionally been provided by computing specialists from the FNAL Computing Division.

Support by the collaboration for code development of the simulation packages has always been problematic and understaffed. With more students and post-docs focusing on physics analysis, it will be challenging to improve the situation in the future. New institutional commitments at MOU level need to be negotiated. In addition, one FTE of a computing professional for projects such as the transition to Geant 4 will be needed in the next couple of years. Furthermore, availability of CPU resources on the CDF central machines for code development and testing needs to continue as well as access to the CAF queue for the validation of the functionality of the cdfSim executable in new offline releases is required.

Overall, the focus in the simulation is shifting from core infrastructure issues to operations support for large MC production, as well as detailed tuning of the Monte Carlo. To fully support this change of focus we have recently combined the CDF MC production group with the simulation group, and have added a third co-leader to focus primarily on MC tuning and validation: the leaders come from CMU, McGill and Duke. The challenge for the new leadership is to motivate more people in the CDF physics analysis groups to take ownership of the physics content in the simulation. By giving the simulation group full control over substantial offsite computing resources, we are hoping to provide an additional carrot to lure the physics analysis focused parts of the collaboration into more actively participating in the work of the simulation group.

15 Validation

15.1 Data Validation

At present, the validation of the physics runs taken by CDF is carried out at two different levels. On-line, a small fraction of the events, passing the trigger decisions, is analyzed by C++ consumer programs, using a common framework installed on B0 control-room computers. Those modules mainly fill histograms that are compared to reference values by the shift crew, who then performs a first on-line decision about the quality of the data. Online histograms monitor the performance of the detectors in terms of occupancies, rates, etc.., as well as the proper reconstruction of high-level physics objects like, for example, jets, tracking, electrons or missing transverse energy. Decisions are finally collected by the shift crew in a reduced set of bits and save into databases. For those runs passing a minimum set of on-line requirements, events are processed off-line using a given version of *production* executable which includes the best knowledge of fundamental parameters like, for example, calibration constants, hot/dead channels or beam alignments, and reconstruct the basic high-level objects used for the physics analyses. In addition, off-line executable also includes a number of validation modules used to reconstruct, for each run, relevant distributions for the different objects and save them into histograms. The output of the validation modules is monitored by member of the CDF off-line validation group, who assign off-line good/bad run bits for different components.

The experience from the past years of CDF operation indicates that a more automated, intense, centralized and coherent effort in on-line data-quality monitoring and off-line physics validation would be desirable. Recently, an experimental group from the High Energy Physics Institute of the University of Barcelona (IFAE) has joined the CDF Collaboration, with an institutional responsibility related to Data Quality Management (DQM). In the following months we expect that on-line decisions by the shift crew in terms of data quality will be automated, supported and supervised by dedicated DQM personnel. An expert system is being built to allow an automatic check of off-line histograms which would make possible a day-by-day based validation of the data as being collected by the experiment. In addition, recent developments in the database structure (and related software) will allow to use run-sections within the run in order to classify the quality of the data, thus increasing the amount of integrated luminosity available for physics analyses.

15.2 Simulation Validation

The GEANT-based CDF detector simulation described in a previous section gives quite a good description of the detector response. The simulation is being used extensively for Run 2 analyses in electroweak, top and B physics, searches for new phenomena and QCD studies.

The geometry of the active detectors - tracking, calorimetry and muon systems, has been adequately simulated and validated. The uncertainty in the predicted lepton acceptances are on the order of 1% of themselves. The spatial distributions of the leptons in W and Z boson decays, for instance, are reproduced by the simulation.

The underlying event generator of PYTHIA has been tuned and reproduces the primary track multiplicity and p_T spectrum in the central outer tracker (COT). The hit resolution in the COT is based on the observed hit residuals of isolated, high p_T tracks. The predicted track parameter resolutions agree with the data within about 10%.

The simulation of the electromagnetic calorimeter (and its component detectors) predicts the electron identification efficiency correctly within about 2%. The energy scale and resolution is modelled adequately, with fine-tuning usually performed at the analysis level. Tools to simulate additional interactions have been validated and are in use to obtain good agreement with data, especially in the plug calorimeter where the electron identification variables are sensitive to the extra energy flow.

A number of different strategies have been pursued for the simulation of charge deposition in the silicon detector. In the future we plan the default technique to be a detailed hit-level parametric model, for reasons of speed and ease of tuning. The code is ready for large-scale usage and will be used for forthcoming simulation runs.

A thorough accounting and implementation of the passive material in the detector has been performed in the simulation geometry. This level of detail is needed for accurate energy loss corrections for tracking, and simulation of bremsstrahlung and conversions for electrons/photons. Precision measurements in electroweak and B physics are driving further improvements in the passive material geometry, using comparisons of the simulation with calibration data ($W, Z \rightarrow e$, $J/\psi \rightarrow \mu\mu, ee$, $\Upsilon \rightarrow \mu\mu$, $K_S \rightarrow \pi\pi$ and cosmic rays). The simulation of secondary interactions of hadrons, which affects the total track multiplicity and potentially affects the response to hadronic jets, is also connected with the passive material geometry. The simulation agrees with the data within 10-30% for these various energy loss mechanisms. Work continues on the implementation and validation of additional details regarding the passive material, with a goal of achieving agreement with the data at the 5% level.

The largest continuing effort is in tuning the hadronic calorimeter response, which is crucial for reproducing jet properties for top and QCD physics. The methods of balancing p_T in dijet and γ-jet events, and calibrating single isolated particle response using the tracker, are being employed. The single particle response in the central region is being cross-checked and improved. The analysis of the plug calorimeter hadronic response is on-going but will need more time, as the tracking is the forward region is not as robust as in the central region. Two new tracking algorithms, which integrate the silicon, COT and calorimetry information in the forward region, have been developed recently. These algorithms are being integrated into the current release of the tracking code, and will be exploited to improve the constraints on the single particle response in the plug. We plan to make significant progress on the plug hadronic tuning over the next year.

16 Code Management

We have considerable experience distributing the CDF offline analysis code on a wide variety of hardware and software platforms, and in a number of runtime environments .

- Offline interactive and batch
- Online Level 3 triggering
- Online monitoring

16.1 Management

We have been running Code Management operations with roughly 3 FTE of effort from the CDF department in the computing division. Recently, this has scaled back to about 2 FTE, as the code and its infrastructure have become a bit more stable, and as we have shifted some of this team's efforts into Data Handing, and SAM in particular.

Primary areas of responsibility are

- Releases - building and packaging releases as needed.

  The tools for this have become much more mature recently, as proven by the ability of several untrained managers to tag, build and distribute releases without direct assistance from Code Management.

- Validation - verifying that a standard set of program build, and run without crashing on standard data files. This is not a full physics certification, but is a useful first step before releasing code to the full collaboration for physics certification.

- Distribution - packaging and installation releases on Fermilab central systems. Supporting and trouble shooting installation on systems worldwide.

16.2 Systems

We operate a few dedicated systems in support of the above activities.

- CDFPCA - Linux code server

  This is the primary code management system, performing many functions as enumerated below.

  1. Build releases
  2. CVS server for CDF and ZOOM
  3. UPD and code distribution server for CDF
  4. CDF and SAM code browser, using lxr software
  5. CDF code management general purpose web server

- CDFPCB - Former Linux code server

  We keep the previous generation CDFPCA hardware around for development and testing of the various servers, as well as special code tests, such as daily clean rebuild of the code, which would put too much load on CDFPCA.

- FCDFCODE1 and NCDF209 - code servers

  We provide centrally managed NFS served copies of all requested offline release, This minimizes the number of onsite systems needing separate installations, and provides performance comparable to locally mounted disks. FCDFCODE1 serves the central CAF and interactive nodes. NCDF209 serves the CDF desktop systems.

16.3 Plans

Before describing existing operations in more detail, let's cut the the bottom line and outline ongoing support needs. So far we have spent about $50K every year for new server systems.

- CDFPCA needs to be upgraded every 2 to 3 years. The external disks have just been upgraded (June 2003). We need a new host system to carry us a few more years. The old 700 MHz PIII processors in CDFPCA will not be sufficient in the future.
- Various licensed software needs to be renewed or acquired. This presently includes KCC, Purify, Insure, and Totalview. The exact list may change as we move from KCC to gcc, and as we evaluate new platforms.
- If we wanted to start purchasing IA64 systems for evaluation, we would need to benchmark these systems with CDF software, both for performance reasons and because we could not purchase these systems until we knew the CDF software could be ported to them. Porting run 2 software to IA64 is a major effort for both CDF and D0, and should be led by a computing division task force.
- Transitions - there is a long list of software transitions and projects needing work, at http://cdfkits.fnal.gov/Dist/doc/transitions.txt We expect the existing personnel to be quite busy with this or similar work for the foreseeable future.
- Continued deployment and support of CDF code to the Grid, for example within the Open Science Grid initiative, requires additional Code Management support, at the level of a short term 2 FTE-years of effort. Extending the scope of code management to include packaging of SAM, or other computing infrastructure software would require an additional FTE on a sustained basis.

16.4 Releases

- Base - major validated and certified releases (4.1.0)
- Point - minor bug fix and operational patches to Base releases (4.1.1)
- Integration - built biweekly as stable base for development, unvalidated
- Development - built daily from CVS head to catch problems early
- Development lite - updated daily from CVS on all systems to provide current management scripts, database configurations and other operations configuration files.
- Patch - test releases built from a simple patch file in CVS. Online Level3, Online Consumer, and Farm Production programs are patch releases.

16.5 Configurations

We try to minimize the number of different configurations supported. The computing world is quite dynamic, and some variations are inevitable. Wherever possible, we make the extra effort to keep the code itself independent of these variations.

- Operating Systems (Fermi Linux 7 and IRIX)

  We have dropped the formerly supported AIX, OSF1 and SunOS, and will drop IRIX soon. We may need to move to Linux 9 soon. We may be evaluating IA64 Linux systems soon.

- Optimizations (debug, C++ inlining and maximum)

  Some packages or routines fail to optimize properly (usually detected as unacceptably large compile times), and must have optimization individually adjusted.

- Databases

  We generate separate calibration libraries with interfaces to text, mSQL, Oracle, MySQL, and ODBC databases.

  We are reducing this to text plus one ODBC type API to all the back end databases, to increase flexibility and maintainability.

- Compilers

  We will drop KCC, moving to gcc 3 during 2003. KCC is no longer supported by the vendor, and it is only a matter of time until we must move to a platform or operating system release on which it does not work.

  We will evaluate the Intel C++ compiler for best performance on Intel processors.

- Kernels

  There are about 21 different Linux kernels in use offsite, each used on only 1 to 4 nodes. Therefore we try very hard to keep the CDF offline code independent of the kernel.

16.6 Product versions

Our releases depend on many relatively stable external products, which we do not build as part of the release. These are distributed and accessed with the UPD/UPS tools, which are essential for this task. Presently, in August 2003, we use

- 21 Computing Division supported products
- 4 CDF supported products

The Computing Division has had to build up to 15 varieties of each release of root, due to variations in platform, kernel, compilers, optimization, and exception handline. We have worked hard to reduce this work load, by moving toward a common gcc compiler, enabling C++ exception handling, and dropping older kernels and platforms.

The development and current product versions are updated by requiring each system to run a nightly update of a 'development lite' release of the CDF code, containing a few standard maintenance and operations scripts, database access configurations, other small operations data files.

17 L3 Code

The Level-3 filter group develops and makes available the executables that are run on the Level-3 farm, and maintains the infrastructure for preparing the calibrations at the start of each run. Further tasks include various book-keeping and archival operations such that Level-3 executables can always be re-created and re-run in the future, and validation of Level-3 specific code. Almost 400 “tagsets” (combinations of executables and control files) have so far been built in Run 2, and over 10,000 calibration tar-balls have been automatically generated. The turn-around time for the creation of a new Level-3 executable is typically a few hours from the time that the new physics table information is entered into the trigger database. Recent additions and improvements to the code running at Level-3 include silicon reconstruction (outside-in tracking only), the writing of Level-3 summary information (enabling the bulk of Level-3 reconstruction objects to be dropped on output) and the compression of the tracking detector raw data banks.

Currently, reconstruction and filtering take approximately 0.7 GHz seconds per event at Level-3. There is a roughly linear dependence on the instantaneous luminosity, due to increased detector occupancy and the different event composition at higher rates. At $\mathcal{L} = 10^{32}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$, the CPU required should still be < 1.2 GHz seconds per event. Additional reconstruction and compression might add $10 - 20\%$ to this. With a total Level-3 farm capacity of 700 GHz (after the forthcoming addition of 64 2 GHz dual nodes), Level-2 accept rates close to 500 Hz could therefore be tolerated before deadtime is incurred.

Significant work has been undertaken recently to reduce the size of the events written out by Level-3. The two main lines of attack have been to drop as many Level-3 reconstruction results as possible without having an adverse impact on physics analysis and monitoring, and to run lossless compression algorithms which enable the uncompressed raw data banks to be dropped. Dropping the Level-3 reconstruction objects results in a saving of $\approx 35-40$ kB per event. Silicon RAW banks have a compression factor of ≈ 3 and the COT banks ≈ 2. The combined effect is an event size in the range $120-130$ kB for events that do not go to Streams A or D (neither uncompressed raw data banks or Level-3 reconstruction objects are dropped for such events). With a Stream A/D event size of $\approx$ 290 kB and an overlap of around 15% with other streams, the average event size is in the region of 150 kB. Compression of remaining raw data banks might reduce this by an additional $10-20$ kB per event. There is some scope for further reductions by utilizing ROOT compression at Level-3, although this would require some development work and the compression factors will presumably be much smaller for data that has already had dedicated compression code run on it.

In terms of computing requirements, the activities of the Level-3 filter group are modest. We heavily utilize a single build node for most executable development work and building operations. The same node runs the calibration generation scripts at run-start. The storage requirements for archival of Level-3 executables and calibrations are very small, < 100 GB/yr. The key requirements of stability and security are identical to all other machines directly involved in online operations.

18 Offline Operations

Offline operations are generally concerned with the day-to-day integration of the operations of code management, critical path code development for new reconstruction executables, online operations, offline calibrations, reconstruction farms, data handling, databases and the central analysis facilities. Offline operations and the offline shift specifically pay close attention to the process of producing, testing and releasing an executable that can be run on the reconstruction farms. The offline shift is one person from the collaboration available 8-4 in a 1 week shift with 2 days of overlap with the prior shift and the next shift. Although they can help with well understood tasks, their expertise for debugging problems and coordinating resolution is limited. From FY01 through FY03 effort in offline operations has primarily come from two people in the CDF department of the computing division. Beginning in August 2003 CDF has instituted a new position, the offline production coordinator, which is a 6 month shift of two people skilled in the offline who can train and supervise the offline shift and coordinate the resolution of problems in the reconstruction executable and other areas of offline operations. They are the offline equivalent of the CDF detector operations managers. We expect this new shift role will enhance the efficiency of offline operations.

19 Summary and Conclusions

19.1 Cost Summary

FY	Batch CPU ($M)	Inter. CPU ($M)	Farm CPU ($M)	DB ($M)	Tape Robot ($M)	Cache Disk ($M)	Net-work ($M)	Legacy Sys. ($M)	Total ($M)
01 A	-	-	0.25	-	-	-	-	0.75	1.0
02 A	0.39	0.07	0.22	0.02	0.77	0.63	0.25	0.69	3.0
03 A	0.31	0.08	0.13	0.14	0.20	0.34	0.23	0.00	1.4
04 E	0.76	0.10	0.19	0.10	0.27	0.20	0.25	-	1.9
05 E	1.16	0.10	0.19	0.10	0.63	0.64	0.19	-	3.0
06 E	0.70	0.10	0.19	0.10	0.39	0.32	0.12	-	1.9

Table 24: CDF computing equipment purchasing plan. The fiscal year, batch CPU for the CAF, Linux based interactive CPU and its local disk, production farm CPU, databases, Enstore tape robot & tape drives, network attached cache disk, networking, legacy SMP CPU and disk systems, and total procurements.

In Table 24 we summarize the actual equipment spending costs for CDF computing by Fermilab in fiscal years 2001 through 2003 and the estimated costs for following our computing plan to meet our requirements for fiscal years 2004 through 2006. The CDF computing budget from Fermilab for 2001 was $2M of which we spent $1M after all purchases and returns were accounted for. In FY02 we spent all of the $2M budget and the $1M not spent in FY01 that was rolled over to FY02. In FY03 the budget was $1.5M and we spent $1.43M on equipment and $0.05M was contributed to upgrade the power in FCC to accommodate the equipment. The lab guidance for FY04-FY06 for the CDF computing equipment budget is $2M per year. Table 24 shows that this should meet roughly 90% of our needs over the three year period.

Table 24 reveals some basic features about computing planning and budgets at CDF. In FY01 expenditures were low as CDF had its first taste of data and a measure of the reality of run 2 computing, and transitioned its computing planning from the SMP and local storage model favored by the run 1 inspired planning to a more network based model utilizing commodity computing. In FY02 the introduction and significant buildup of the CAF and the network attached tape drives produced a surge in spending. In FY02 relatively modest CPU expenditures produced a large increase in CPU capacity by early FY03. In FY03 the reduced Fermilab budget of \$1.5M did not allow a significant increase in CPU expenditures. The significant increase in spending on CPU in FY04 reflects the need to analyze large datasets that will occupy an increasing part of the CDF data volume with the reduction in the event size that occurs in FY04. If the lab provides \$2M per year in FY04 through FY06, there is a significant funding shortfall in FY05 when the bandwidth capability for raw data logging is anticipated to double and our storage and CPU costs increase. The 50% increase in the online data logging rate in FY06 for the run 2B upgrade increases our needs for tape drives to sustain peak loads, but the shutdown decreases the data logged and our disk and CPU requirements are reduced.

If the lab guidance for FY04-FY06 should shrink to \$1.5M per year, the amount budgeted and spent in FY03, then the CDF batch CPU requirements will not be met by Fermilab. This would produce a challenging offline environment where CDF will need to rely significantly on the resources of offsite institutions to satisfy its batch analysis requirements. Although offsite institutions already provide the majority of our Monte Carlo production needs, the GRID technology to allow them to contribute to global CDF user analysis and global data handling requirements has not been implemented yet at CDF. There is only a small chance that offsite institutions can contribute significantly to the global analysis needs of CDF users as early as FY04. Our plan is that GRID technology should be available for first production level use in FY05, and realistically we should not rely formally on offsite contributions to our global analysis needs until FY06.

The costs in Table 24 do not include operating costs. Table 16 showed that we anticipate no operating costs from tapes in FY04 due to our policy of recycling old tapes. In FY05 the doubling of the raw data logging capability, and the migration of tapes to a new media, result in tape expenditures of \$0.35M, which decrease to \$0.14M in FY06 due to the shutdown. In addition to these tape costs, as in previous years we anticipate roughly \$0.2M in additional operating costs each fiscal year from FY04-FY06 due to miscellaneous expenditures on installations, cabling, racks, FNAL desktops, consultants, etc.

19.2 Operations Summary

We have commissioned and operated an offline that has reconstructed and analyzed 0.5 billion unique events of raw data and wrote 0.5 PB of data to tape. The CAF has demonstrated that we can successfully utilize cheap commodity computing for our analysis needs. Enstore and the CAF have demonstrated that the networking for distributed computing is robust and reliable as we enter an era where distributed computing is increasingly important. We are operating a data handling system based on dCache and Enstore that consistently delivers 20 TB/day for user analysis, and we have made progress in migrating to SAM. The replicated database systems at CDF are handling an increasing load of connections from multiple computers on the CAF and an increased level of remote connections. We have successfully transitioned our Monte Carlo production activities from the reconstruction farms to offsite institutions. We have made significant progress in all areas of offline software. The results for the summer 2003 conferences included as much as 0.2 fb^{-1} of data and were compared to the predictions of a full CDF simulation. While these achievements indicate that CDF computing and the offline are functioning, we realize that we have recorded only about

3% of the total luminosity projected from Run 2, and most of the challenge is in front of us. For the existing systems, the entire previous three fiscal years could be viewed as an evolution and commissioning phase, and if we are fortunate the next three will be a phase of optimization and operations for these systems. While scaling the Fermilab systems provides the maximum operational and budgetary efficiency overall, the Fermilab budget may not be able to sustain the eventual equipment needs or support the total required operational personnel. The technology and financing in computing today are moving towards a distributed model (GRID), and CDF is preparing to take advantage of it.

References

[1] CDF Plan and Budget for Computing in Run 2, Edited by Robert M. Harris, with contributions from W. Badgett, S. Belforte, P. Demar, R. Jetton, K. McFarland, D. Petravick, D. Tang, J. Tseng, S. Wolbers and F. Würthwein, CDF Note 5914, May 2002.

[2] R. Snider and R. Harris, "Implications of increased data logging rate on the CDF Run 2 computing plan and budget", CDF Note 6639, August 21, 2003.

[3] The Run II Luminosity Upgrade at the Fermilab Tevatron, Project Plan and Resource-Loaded Schedule, June 15, 2003.

[4] CDF-5785, "Physics Analysis Computing needs assessment, CDF CAF Review, Fall 2001", 2001.

[5] CDF-5802, "Final Report, CDF CAF Review, Fall 2001", 2001.

[6] Frank Wurthwein, "User Analysis Computing at CDF", presentation at the CHEP 2003 conference, 2003.

[7] I. Bird *et al*, Report of the Run 2 Computing Review, June 4-6, 2002.

[8] R. Brun and F. Carminati, "GEANT Detector Description and Simulation Tool", CERN Programming Library Long Writeup **W5013** (1993)